¿QUÉ ES LA MATEMÁTICA?

RICHARD COURANT
Jefe del Departamento de Matemáticas
de la Universidad de Nueva York

HERBERT ROBBINS
Profesor de Matemáticas
de la Universidad de Nueva York

¿QUÉ ES LA MATEMÁTICA?

UNA EXPOSICIÓN ELEMENTAL DE SUS IDEAS Y MÉTODOS

Traducción del inglés por
LUIS BRAVO GALA
Licenciado en Ciencias Exactas



AGUILAR

colección ciencia y técnica
sección matemáticas y estadística
obra incorporada con el asesoramiento
de luis bravo gala

edición española

depósito legal m 14137/1979
quinta edición–segunda reimpresión–1979
ISBN 84-03-20032-3
printed in spain impreso en españa por gráficas halar s l
andrés de la cuerda 4 madrid

edición original
what is mathematics?
oxford university press new york and london

PRÓLOGOS

PRÓLOGO A LA PRIMERA EDICIÓN

Desde hace más de dos milenios, una cierta familiaridad con la matemática ha sido considerada como parte indispensable de la formación intelectual de toda persona cultivada. En la actualidad, sin embargo, se halla en grave peligro el puesto ocupado tradicionalmente en la educación por esta disciplina; por desgracia, algunos de los profesionales que la representan comparten la responsabilidad de tal situación. La enseñanza de la matemática ha degenerado con frecuencia en un vacío entrenamiento de resolución de problemas, que si bien puede desarrollar una habilidad formal, no conduce en cambio a una comprensión efectiva ni a una mayor independencia intelectual. La investigación matemática muestra una tendencia hacia la superespecialización y hacia una excesiva insistencia en lo abstracto; las aplicaciones y conexiones con otros campos del saber han sido descuidadas. Sin embargo, tal estado de cosas no debe justificar una política de retraimiento. Por el contrario, la reacción opuesta puede y debe partir de aquellos que se sienten conscientes del valor de la disciplina intelectual. Profesores, estudiantes y público culto piden una reforma constructiva y no una resignación siguiendo la línea de menor resistencia. La meta será una verdadera comprensión de la matemática como un todo orgánico y como base para el pensamiento y la acción científicos.

Algunos libros espléndidos de biografía e historia y otros más populares han estimulado el interés general latente; pero el conocimiento no puede adquirirse utilizando únicamente medios indirectos. La comprensión de la matemática no puede ser transmitida indirectamente o como un juego sin dificultades, como tampoco pueden adquirir una educación musical, a través de reseñas periodísticas brillantes, aquellos que no han escuchado buena música con frecuencia. Un contacto real con el *contenido* de la matemática viva es necesario. Sin embargo, cabe evitar algunos detalles de técnica y muchas digresiones; la presentación de las matemáticas debe estar tan libre de excesos de rutina como del dogmatismo prohibitivo que rehuye revelar el motivo o la meta y que constituye así un obstáculo de

de mala fe para un esfuerzo honesto. Es posible seguir una ruta directa a partir de los elementos fundamentales hasta puntos avanzados, desde los cuales puedan divisarse la sustancia y las fuerzas directrices de la matemática moderna.

El presente libro es un intento en esa dirección. Y en tanto que presupone únicamente los conocimientos que pueden adquirirse en la enseñanza media, puede considerarse como elemental. No constituye, sin embargo, una concesión a la tendencia peligrosa que consiste en soslayar toda dificultad. Su lectura requiere un cierto grado de madurez intelectual y un deseo de tener pensamientos propios. El libro está escrito para principiantes y entendidos, para estudiantes y profesores, para filósofos e ingenieros, como libro de texto y de consulta. Quizá esto encierre una intención demasiado ambiciosa. Bajo la presión de otros trabajos han sido hechas ciertas concesiones al publicar el libro, después de varios años de preparación, antes de estar realmente terminado. Por ello, críticas y sugestiones serán bien recibidas.

De todos modos, se espera que el libro pueda servir un propósito útil y constituir una contribución a la educación superior americana, procedente de alguien que está profundamente agradecido por la oportunidad que le fué ofrecida en este país. Mientras que la responsabilidad por el plan y la filosofía de esta publicación recae enteramente en el que suscribe, todos los méritos que pueda encerrar debe compartirlos con Herbert Robbins. Éste, desde el momento en que fué asociado a esta tarea, la ha considerado generosamente como su propia causa, y su colaboración ha desempeñado un papel decisivo en el trabajo de completar la obra en su forma presente.

Debo reconocimiento también a la ayuda de varios amigos. Discusiones con Niels Bohr, Kurt Friedrichs y Otto Neugebauer han influído en la actitud filosófica e histórica; Edna Kramer nos ha hecho críticas constructivas desde el punto de vista de la enseñanza; David Gilbarg preparó las primeras notas de curso que originaron el libro; Ernest Courant, Norman Davids, Charles de Prima, Alfred Horn, Herbert Mintzer, Wolfgang Wasow y otros prestaron su ayuda en la interminable tarea de escribir varias veces el manuscrito, y contribuyeron a muchas mejoras de detalle; Donald Flanders hizo varias sugestiones valiosas y arregló el manuscrito para la imprenta; John

Knudsen, Hertha von Gumppenberg, Irving Ritter y Otto Neugebauer prepararon las figuras; H. Whitney contribuyó a la colección de ejercicios del Apéndice. La General Education Board de la Fundación Rockefeller ha contribuído generosamente al desarrollo de los cursos y notas que fueron luego la base del libro. Gracias sean dadas también a la Waverly Press, y en particular a Grover C. Orth, por su trabajo, de una competencia extraordinaria; y a la Oxford University Press, en especial a Philip Vaudrin y W. Oman, por su animosa iniciativa y cooperación.

R. COURANT.

NEW ROCHELLE, N. Y.
Agosto de 1941.

PRÓLOGO A LA SEGUNDA, TERCERA Y CUARTA EDICIONES

Durante los últimos años, la fuerza de los acontecimientos condujo a una creciente demanda de información y enseñanza matemáticas. Ahora más que nunca existe el peligro de frustración y desilusión, a no ser que estudiantes y profesores intenten ver más allá del formalismo y manipulación matemáticos, y comprendan la verdadera esencia de la matemática. Este libro fué escrito para esos profesores y estudiantes, y la acogida hecha a la primera edición confirma a los autores en la esperanza de que pueda ser de utilidad.

Críticas de varios lectores han dado lugar a numerosas correcciones y mejoras. Gracias cordiales debemos a Natascha Artin por su generosa ayuda en la preparación de la cuarta edición.

R. COURANT.

NEW ROCHELLE, N. Y.
18 de marzo de 1943.
10 de octubre de 1945.
28 de octubre de 1947.

CÓMO DEBE UTILIZARSE ESTE LIBRO

El libro está escrito en un orden sistemático, pero esto no quiere decir que sea necesario que el lector lo estudie página a página y capítulo tras capítulo. Por ejemplo, la introducción histórica y filosófica puede muy bien dejarse para después de haber leído el resto. Los diferentes capítulos son, en gran parte, independientes unos de otros. Frecuentemente, el comienzo de una sección podrá ser comprendido sin dificultad. Luego, el camino conducirá gradualmente hacia adelante, para llegar en el final de cada capítulo y en los suplementos a cuestiones más difíciles. Así, el lector que desee información general más bien que conocimientos específicos puede conformarse con una selección de materias que eluda los análisis detallados.

El estudiante con poca base matemática también puede hacer una selección. Asteriscos y tipos pequeños de letra indican partes que pueden ser omitidas en una primera lectura sin gran perjuicio para la comprensión de las siguientes. Por otra parte, no habrá inconveniente si el estudio se limita a las secciones o capítulos en los que el lector esté más interesado. La mayoría de los ejercicios son de carácter no rutinario; los de mayor dificultad van marcados con un asterisco. El lector no debe alarmarse si no alcanza a resolver algunos de ellos.

En los capítulos sobre construcciones geométricas y sobre máximos y mínimos se encontrará material apropiado para grupos selectos de estudiantes de cursos superiores.

Esperamos que el libro servirá tanto para el estudiante que comienza como para el que esté más avanzado, así como para el profesional que se halle verdaderamente interesado en la matemática. Además, puede servir como base para cursos de tipo no tradicional sobre los conceptos fundamentales de esta ciencia. Los capítulos III, IV y V pueden utilizarse en un curso de geometría, mientras los capítulos VI y VIII ofrecen en conjunto una exposición autónoma del cálculo infinitesimal, en la que se concede atención especial a los conceptos básicos y se trata de evitar los métodos rutinarios. Pueden usarse como un texto inicial por un profesor que desee hacer contri-

buciones propias, completando el material de acuerdo con sus necesidades específicas y, especialmente, añadiendo otros ejemplos numéricos. Numerosos ejercicios distribuídos a lo largo del texto y la colección adicional del final facilitarán el uso del libro en las clases.

Esperamos también que el lector encuentre a menudo detalles de interés en muchas discusiones elementales que contienen el germen de más amplios desarróllos.

ÍNDICE GENERAL

ÍNDICE GENERAL

¿QUÉ ES LA MATEMÁTICA?

INTRODUCCIÓN

¿QUÉ ES LA MATEMÁTICA?

La matemática, como una expresión de la mente humana, refleja la voluntad activa, la razón contemplativa y el deseo de perfección estética. Sus elementos básicos son: lógica e intuición, análisis y construcción, generalidad y particularidad. Aunque diversas tradiciones han destacado aspectos diferentes, es únicamente el juego de estas fuerzas opuestas y la lucha por su síntesis lo que constituye la vida, la utilidad y el supremo valor de la ciencia matemática.

Sin duda, todo el desarrollo matemático ha tenido sus raíces psicológicas en necesidades más o menos prácticas. Pero una vez en marcha, bajo la presión de las aplicaciones necesarias, dicho desarrollo gana impulso en sí mismo y trasciende los confines de una utilidad inmediata. Esta tendencia de la ciencia aplicada hacia la teórica aparece tanto en la historia antigua como en muchas de las contribuciones a la matemática moderna debidas a ingenieros y físicos.

La historia de las matemáticas comienza en Oriente, donde, hacia el año 2000 a. de J.C., los babilonios poseían ya una gran cantidad de material que podría ser clasificado hoy como perteneciente al álgebra elemental. Pero como ciencia, en el sentido moderno, la matemática aparece más tarde, en Grecia, entre los siglos v y IV antes de J.C. El contacto creciente entre el Oriente y los griegos, que comienza en los tiempos del imperio persa y culmina en el período que sigue a las expediciones de Alejandro, puso a los griegos al corriente de los conocimientos de los babilonios en matemática y astronomía. La matemática fué sometida entonces a las discusiones filosóficas que florecieron en las ciudades griegas. Los pensadores griegos se dieron pronto cuenta de las grandes dificultades inherentes a los conceptos matemáticos de continuidad, movimiento e infinitud, así como al problema de medir magnitudes arbitrarias con unidades prefijadas. Entonces fué llevado a cabo un admirable esfuerzo para vencerlas y el resultado, la teoría de Eudoxio del continuo geométrico, fué de tal perfección, que para encontrar algo que pueda comparársele es necesario qué, dos milenios más tarde, aparezca la teoría moderna de los números irracionales. La tendencia axiomático-deductiva en matemáticas tuvo su origen en tiempos de Eudoxio y cristalizó en los *Elementos* de Euclides.

Sin embargo, aunque la tendencia teórica y axiomática de la matemática griega es una de sus más importantes características y ha ejercido una influencia enorme, nunca se insistirá demasiado en que las aplicaciones y conexiones con la realidad física desempeñaron un papel importante como parte de la matemática de la antigüedad, y que en muchas ocasiones fué preferido un modo de exposición menos rígido que el de Euclides.

Es muy posible que el descubrimiento de las dificultades relacionadas con las cantidades *inconmensurables* desviara a los griegos del desarrollo del cálculo numérico, alcanzado con anterioridad en Oriente. En su lugar, se abrieron camino a través de la geometría axiomática pura. Y así comenzó un extraño rodeo en la historia de la ciencia, y quizá se perdió una gran oportunidad. Durante casi dos mil años el peso de la tradición geométrica griega retrasó la inevitable evolución del concepto de número y el desarrollo del cálculo algebraico, que más tarde habían de ser la base de la ciencia moderna.

Después de un período de preparación lenta, la revolución en la matemática y en la ciencia comenzó su fase vigorosa en el siglo XVII, con la geometría analítica y el cálculo diferencial e integral. Mientras la geometría griega conserva aún un lugar destacado, el ideal griego de cristalización axiomática y de deducción sistemática desaparece durante los siglos XVII y XVIII. Razonamientos lógicos rigurosos, a partir de definiciones claras y no contradictorias, axiomas *evidentes*, fueron cuestiones sin importancia para los nuevos exploradores de la ciencia matemática. En una verdadera orgía de conjeturas intuitivas, de razonamientos convincentes entrelazados con un misticismo sin sentido, con una confianza ciega en el poder sobrehumano de los procesos formales, conquistaron un mundo matemático de inmensas riquezas. Luego, gradualmente, la exaltación del progreso dejó el paso a un espíritu de autocrítica. En el siglo XIX la necesidad inmanente de consolidar, y el deseo de una mayor seguridad en la extensión de la enseñanza superior, que había impulsado la Revolución francesa, condujo inevitablemente a una revisión de los fundamentos de la nueva matemática, en particular del cálculo diferencial e integral, así como del concepto fundamental de límite. Así, el siglo XIX constituyó no sólo un período de nuevos avances, sino que además puede caracterizarse por un afortunado retorno al ideal clásico de precisión y demostraciones rigurosas. Y en este sentido llegó a superar al modelo de la ciencia griega. Una vez más el péndulo se inclinó del lado de la pureza lógica y de la abstracción. Actualmente vivimos aún en este período, aunque es de esperar que la desafortunada se-

paración entre la matemática pura y las aplicaciones a la vida, quizá inevitable en tiempos de revisión crítica, venga seguida de una era de íntima unidad.

La renovada solidez interna, y sobre todo la simplificación enorme alcanzada sobre la base de una comprensión más clara, hacen posible hoy poder dominar la teoría matemática sin perder de vista las aplicaciones. Establecer de nuevo una unión orgánica entre ciencia pura y aplicada y un equilibrio estable entre la generalidad abstracta y la individualidad concreta puede ser muy bien la tarea universal de la matemática en el futuro inmediato.

No es éste el lugar para un análisis filosófico o psicológico detallado de la matemática. Únicamente podemos destacar algunos puntos. Parece existir un grave peligro en el excesivo predominio del carácter axiomático-deductivo de las matemáticas. Ciertamente, el elemento de invención constructiva, de intuición directora, escapa a una simple formulación filosófica; sin embargo, continúa siendo el núcleo de todo resultado matemático, aun en los campos más abstractos. Si la forma deductiva cristalizada es la meta, la intuición y la construcción son, cuando menos, las fuerzas directrices. Una amenaza seria para la verdadera vida de la ciencia aparece contenida en la afirmación de que la matemática no es más que un sistema de conclusiones derivadas de definiciones y postulados que deben ser compatibles, pero que, por lo demás, pueden ser creación de la libre voluntad del matemático. Si esta descripción fuera exacta, las matemáticas no podrían interesar a ninguna persona inteligente. Sería un juego con definiciones, reglas y silogismos, sin meta ni motivo alguno. La noción de que el intelecto puede crear sistemas de postulados plenos de significado de modo arbitrario es una verdad «a medias» decepcionante. Únicamente bajo una disciplina de responsabilidad frente a un todo orgánico, guiada sólo por necesidades intrínsecas, puede la mente libre obtener resultados de valor científico.

Aunque la tendencia pasiva del análisis lógico no puede representar toda la matemática, ha conducido, sin embargo, a una comprensión más profunda de los hechos matemáticos y de su interdependencia, y también a una mayor penetración en la esencia de los conceptos matemáticos. A partir de ella se ha desarrollado un punto de vista moderno en las matemáticas que es característico de una actitud científica universal.

Cualquiera que sea el punto de vista filosófico, para todos los propósitos de observación científica, un objeto agota en sí la totalidad de relaciones posibles respecto del observador o del instrumento. Na-

turalmente, la simple percepción no constituye conocimiento; debe ser coordinada e interpretada con referencia a alguna entidad subyacente, una «cosa en sí», que no es un objeto de la observación física directa, sino que pertenece a la metafísica. Sin embargo, en el proceso científico es importante descartar los elementos de carácter metafísico y considerar los hechos observables como la última fuente de nociones y construcciones. Renunciar a la meta de comprender la «cosa en sí», de conocer la «realidad última», de desentrañar la esencia más íntima del mundo, puede ser psicológicamente penoso para entusiastas ingenuos, pero de hecho es uno de los sacrificios de consecuencias más fecundas en el pensamiento moderno.

Algunos de los mayores avances en la física han sido el premio a una adhesión decidida al principio de eliminar la metafísica. Cuando Einstein consiguió reducir la noción de «sucesos simultáneos que ocurren en lugares distintos» a fenómenos observables; cuando señaló como prejuicio metafísico la creencia de que este concepto debe tener un significado científico en sí mismo, encontró la clave de su teoría de la relatividad. Cuando Niels Bohr y sus discípulos analizaron el hecho de que toda observación física va acompañada de un efecto del instrumento observador en el objeto observado, se hizo claro que el intento de fijar simultáneamente la posición y la velocidad de una partícula no es posible en el sentido de la física. Las consecuencias trascendentes de este descubrimiento, contenidas en la teoría moderna de la mecánica cuántica, son hoy familiares a todo físico. En el siglo pasado prevaleció la idea de que las fuerzas mecánicas y los movimientos de partículas en el espacio eran cosas en sí mismas, mientras que electricidad, luz y magnetismo debían ser reducidos o *explicados* como fenómenos mecánicos, de la misma manera que se hacía con el calor. El *éter* fué inventado como medio hipotético capaz de los movimientos mecánicos no explicados satisfactoriamente y que aparecían bajo las formas de luz y electricidad. Poco a poco se comprendió que el éter era necesariamente inobservable; por consiguiente, no pertenecía a la física, sino a la metafísica. Con pena por algunos y con satisfacción por otros, las explicaciones mecánicas de la luz y la electricidad, y con ellas el éter, debieron ser finalmente abandonadas.

Una situación análoga, quizá más acentuada, existe en la matemática. A través de los tiempos, los matemáticos consideraron sus objetos, tales como números, puntos, etc., como cosas sustanciales en sí. Pero en vista de que estos entes desafiaban siempre los intentos para una descripción adecuada, los matemáticos del siglo pasado llegaron paulatinamente a la convicción de que el problema de la signifi-

cación de dichos objetos como cosas sustanciales no tenía, en modo alguno, sentido dentro de las matemáticas. Las únicas proposiciones relativas a ellos que pueden importar no se refieren a su realidad sustancial; representan únicamente las relaciones mutuas entre «objetos indefinidos» y las reglas que rigen las operaciones con ellos. Lo que «realmente» *son* los puntos, las rectas y los números ni se puede ni es necesario discutirlo en la ciencia matemática. Lo que interesa y lo que corresponde a hechos *comprobables* es su estructura y relación: que dos puntos determinan una recta, que los números se combinan según ciertas reglas para formar otros números, etc. La percepción clara de la necesidad de una *desustanciación* de los conceptos elementales matemáticos ha sido uno de los resultados más importantes y fecundos del desarrollo axiomático moderno.

Por suerte, las mentes creadoras olvidan las creencias filosóficas dogmáticas cuando la persistencia en ellas podría impedir resultados constructivos. Tanto para entendidos como para profanos no es la filosofía, y sí únicamente la experiencia activa en matemáticas, la que puede responder a la pregunta: ¿Qué es la matemática?

CAPÍTULO PRIMERO

LOS NÚMEROS NATURALES

Introducción.—Los números son la base de la matemática moderna. Ahora bien: ¿qué es un número? ¿Qué significado tiene decir $\frac{1}{2} + \frac{1}{2} = 1$, $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ y $(-1)(-1) = 1$? En la segunda enseñanza se aprende el manejo de las fracciones y de los números negativos, pero una comprensión efectiva de los sistemas de números requiere investigar sus elementos más simples. Mientras los griegos hacen de los conceptos geométricos de punto y recta la base de sus matemáticas, hoy se admite como principio director que todas las proposiciones matemáticas pueden ser reducidas en última instancia a proposiciones sobre los *números naturales*: 1, 2, 3,... «Dios creó los números naturales; el resto es obra de los hombres.» Con estas palabras Leopold Kronecker (1823-1891) señaló la base precisa sobre la cual puede construirse el edificio de la matemática.

Creados por la mente humana para contar objetos agrupados de diversos modos, los números no contienen referencia alguna a las características de los objetos contados. El número 6 es una abstracción obtenida a partir de todas las colecciones que contienen seis cosas; no depende de las cualidades específicas de dichas cosas ni de los símbolos usados para representarlas. Únicamente en etapas avanzadas del desarrollo intelectual llega a percibirse con toda claridad el carácter abstracto de la idea de número. Para los niños, los números están siempre ligados a objetos tangibles, tales como dedos o bolas, y los lenguajes primitivos dan a los números un sentido concreto, designando con palabras distintas los números que corresponden a diferentes tipos de objetos.

Por suerte, los matemáticos no tienen que ocuparse del aspecto filosófico de la transición que da el paso de colecciones de objetos concretos al concepto abstracto de número.

Consideraremos, por tanto, como dados los números naturales, junto con las dos operaciones fundamentales, adición y multiplicación, mediante las cuales pueden ser combinados.

I. CÁLCULO CON NÚMEROS ENTEROS

1. Leyes de la aritmética.—La teoría matemática de los números naturales o *enteros positivos* se conoce con el nombre de *aritmética*.

Se basa en el hecho de que la adición y multiplicación de enteros están regidas por ciertas leyes. Para enunciar estas leyes con toda generalidad no podemos limitarnos a usar símbolos tales como 1, 2, 3, que se refieren a enteros particulares. La proposición

$$1 + 2 = 2 + 1$$

es únicamente un caso particular de la ley general que dice que la suma de dos enteros no depende del orden en que éstos se consideren. En consecuencia, si queremos expresar el hecho de que una cierta relación entre enteros es válida, cualesquiera que sean los valores de los enteros considerados, debemos representarlos simbólicamente mediante letras: a, b, c,... De acuerdo con esto podemos enunciar las conocidas cinco leyes fundamentales de la aritmética:

1) $a + b = b + a$.
2) $ab = ba$.
3) $a + (b + c) = (a + b) + c$.
4) $a(bc) = (ab)c$.
5) $a(b + c) = ab + ac$.

Las dos primeras son las leyes *conmutativas* de la adición y de la multiplicación; indican que el orden de los elementos que intervienen en dichas operaciones puede ser alterado. La tercera, que es la ley *asociativa* de la adición, dice que, en la adición de tres números, se obtiene el mismo resultado si se añade al primero la suma del segundo y el tercero, o al tercero la suma del primero y el segundo. La cuarta es la ley asociativa de la multiplicación. La última, que es la ley *distributiva*, expresa que para multiplicar una suma por un entero se puede multiplicar cada término de la suma por dicho entero y sumar luego los productos obtenidos.

Estas leyes de la aritmética son muy simples e incluso parecen evidentes. Sin embargo, pueden no ser aplicables a otros entes distintos de los números enteros. Así, si a y b son símbolos no de enteros, sino de sustancias químicas, y si *adición* se entiende en su sentido corriente, es evidente que la ley conmutativa no es siempre válida; p. ej., si se añade ácido sulfúrico a agua se obtiene una solución diluída, mientras que la adición de agua a ácido sulfúrico puro puede acarrear consecuencias catastróficas para el experimentador. Análogas consideraciones probarían que, en este tipo de «aritmética» química, las leyes asociativa y distributiva de la adición pueden muy bien no ser ciertas. Asimismo es probable imaginar tipos de aritméticas en las cuales alguna o varias de las leyes 1) a 5) no sean válidas. De hecho, tales sistemas son estudiados efectivamente en la matemática moderna.

Un modelo concreto para el concepto abstracto de entero indicará bien las bases intuitivas sobre las que reposan las leyes 1) a 5). En

vez de usar los símbolos habituales 1, 2, 3,..., representemos los enteros que dan el número de objetos de una colección determinada (p. ej., colección de manzanas de un árbol particular) mediante puntos situados en casillas rectangulares, de modo que a cada objeto corresponda un punto. Operando con estos rectángulos se pueden estudiar las leyes de la aritmética de los enteros. Para sumar dos enteros a y b, dispongamos los rectángulos correspondientes uno contiguo al otro, suprimiendo la raya de separación.

FIG. 1.—Adición.

Para multiplicar a por b, coloquemos los puntos de los dos rectángulos en filas y formemos un nuevo rectángulo con a filas y b columnas de puntos. Las reglas 1)-5)

FIG. 2.—Multiplicación.

corresponderán entonces de modo intuitivo a propiedades evidentes de las operaciones indicadas con los rectángulos.

FIG. 3.—La ley distributiva.

A partir de la definición de adición de dos enteros se puede definir la relación de *desigualdad*. Las dos proposiciones equivalentes, $a < b$ (léase «a menor que b») y $b > a$ (léase «b mayor que a»), significan que el rectángulo b puede obtenerse a partir del a mediante la adición de un tercer rectángulo conveniente, c, de modo que se tenga $b = a + c$. Si escribimos esto en la forma

$$c = b - a,$$

definimos la operación llamada *sustracción*.

FIG. 4.—Sustracción.

La adición y la sustracción se llaman *operaciones inversas*, ya que si la adición del entero d al entero a es seguida de la sustracción del entero d de la suma obtenida, el resultado es el entero inicial a:

$$(a + d) - d = a.$$

Debe observarse que el entero $b - a$ ha sido definido únicamente en el caso en que se tenga $b > a$. La interpretación del símbolo $b - a$ como *entero negativo* cuando sea $b < a$ será discutida más adelante (páginas 63 y siguientes).

A menudo es conveniente usar una de las notaciones equivalentes, $b \geqslant a$ (léase «b mayor o igual que a») o $a \leqslant b$ (léase «a menor o igual que b»), para indicar la proposición contraria de la $a > b$; p. ej., escribiremos: $2 \geqslant 2$, y $3 \geqslant 2$.

Haremos ahora una pequeña ampliación del dominio de los enteros positivos, representados por rectángulos con puntos, introduciendo el entero *cero*, que representaremos por un rectángulo vacío. Denotando éste con el habitual símbolo 0, se tiene, de acuerdo con nuestra definición de adición y multiplicación:

$$a + 0 = a,$$
$$a \cdot 0 = 0,$$

cualquiera que sea el entero a. En efecto, $a + 0$ indica la adición de un rectángulo vacío al rectángulo a, mientras que $a \cdot 0$ indica un rectángulo sin columnas; es decir, vacío. Parece entonces natural extender también la definición de sustracción, poniendo

$$a - a = 0$$

para todo entero a. Se tienen así las propiedades aritméticas características del cero.

Modelos geométricos análogos a los rectángulos de puntos, tales como los antiguos ábacos, fueron usados frecuentemente hasta fines de la Edad Media; pero quedaron desplazados poco a poco por otros métodos simbólicos más cómodos, basados en el sistema decimal.

2. Representación de los números enteros.—Ha de distinguirse netamente entre un número entero y los símbolos 5, V, ... usados para representarlo. En el sistema decimal los diez símbolos: 0, 1, 2, 3, ..., 9, se utilizan para representar el cero y los nueve primeros enteros positivos. Un entero mayor, p. ej., el «trescientos setenta y dos», puede expresarse en la forma

$$300 + 70 + 2 = 3 \cdot 10^2 + 7 \cdot 10 + 2,$$

y se representa en el sistema decimal por el símbolo 372. Es de fundamental importancia observar que aquí la significación de los símbolos 3, 7, 2 es *relativa* y depende de su *posición* en el lugar de las unidades, decenas o centenas. Con esta «notación relativa» se puede representar cualquier entero utilizando únicamente los símbolos de los diez primeros números en varias combinaciones. La regla general consiste en expresar todo entero en la forma ilustrada por el ejemplo

$$z = a \cdot 10^3 + b \cdot 10^2 + c \cdot 10 + d,$$

donde los números a, b, c, d son enteros de cero a nueve. El entero z se representa entonces en la forma abreviada

$$abcd.$$

Notemos de paso que los coeficientes d, c, b, a son los restos obtenidos mediante divisiones sucesivas de z por 10; así

$$\begin{array}{r|l}
372 & 10 \\ \hline
2 & \begin{array}{r|l} 37 & 10 \\ \hline 7 & \begin{array}{r|l} 3 & 10 \\ \hline 3 & 0 \end{array} \end{array}
\end{array}$$

La expresión particular dada antes para z puede servir únicamente para números menores que 10 000, ya que para enteros mayores serán necesarias más de cuatro cifras. Si z es un entero comprendido entre 10 000 y 100 000, lo podremos escribir en la forma

$$z = a \cdot 10^4 + b \cdot 10^3 + c \cdot 10^2 + d \cdot 10 + e,$$

y representarlo por el símbolo *abcde*. Análogas consideraciones valen para enteros comprendidos entre 100 000 y 1 000 000, etc. Será conveniente disponer de un medio de indicar el resultado con toda generalidad mediante una sola fórmula. Para ello designemos los distintos coeficientes: e, d, c,..., con una sola letra afectada de subíndices: a_0, a_1, a_2, a_3, ..., e indiquemos el hecho de que las potencias de 10 pueden ser tan grandes como sea necesario designando la mayor potencia que interviene, no con 10^3 ó 10^4, como en los ejemplos anteriores, sino con 10^n, donde n ha de interpretarse como un entero arbitrario. Entonces el método general para representar un entero z en el sistema decimal consistirá en expresar z en la forma

$$z = a_n \cdot 10^n + a_{n-1} \cdot 10^{n-1} + \dots + a_1 \cdot 10 + a_0, \qquad [1]$$

y representarlo por el símbolo

$$a_n a_{n-1} a_{n-2} \dots a_1 a_0.$$

Como en los casos anteriores, los números $a_0, a_1, a_2, \dots, a_n$, comprendidos entre 0 y 9, serán precisamente los restos sucesivos de dividir z repetidamente por 10.

En el sistema decimal el número 10 ha sido tomado especialmente para servir de base. A primera vista quizá no resalte el hecho de que la elección del 10 no es esencial, y que cualquier entero mayor que uno habría podido servir para el mismo objeto; p. ej., podría usarse un sistema *septimal* (de base 7). En dicho sistema, un entero se expresaría en la forma

$$b_n \cdot 7^n + b_{n-1} \cdot 7^{n-1} + \dots + b_1 \cdot 7 + b_0, \qquad [2]$$

donde las b serían enteros, elegidos entre 0, 1, ... y 6, y se representaría por el símbolo

$$b_n b_{n-1} \dots b_1 b_0 .$$

Así, «ciento nueve» se representaría en el sistema septimal por el símbolo 214, que significaría

$$2 \cdot 7^2 + 1 \cdot 7 + 4.$$

Como ejercicio, el lector puede probar la regla general que permite pasar de la base diez a otra base cualquiera B, y que consiste en efectuar divisiones sucesivas del número z por B; los restos sucesivos serán las cifras del número en el sistema de base B; p. ej.:

$$\begin{array}{r|l}
109 & 7 \\ \hline
4 & 15 \;|\; 7 \\
 & 1 \;|\; 2 \;|\; 7 \\
 & \quad 2 \;|\; 0
\end{array}$$

109 (en el sistema decimal) = 214 (en el sistema septimal).

Es natural preguntarse: ¿cuál será la base más conveniente? Puede verse que bases muy pequeñas presentan inconvenientes, mientras que una base grande requiere la utilización de muchas cifras distintas y, en particular, una tabla de multiplicación extensa. La elección de doce como base ha sido defendida por muchos; como ventajas señalan que 12 es divisible por 2, 3, 4 y 6, y, como consecuencia, el cálculo que implique divisiones y fracciones se simplificaría frecuentemente. Para escribir un número en base doce (sistema duodecimal) se necesitan dos nuevas cifras para el 10 y el 11. Indiquemos el 10 con α y el 11 con β.

En el sistema duodecimal, «doce» se escribiría 10; «veintidós» sería 1α; «veintitrés» sería 1β, y «ciento treinta y uno» se escribiría $\alpha\beta$.

La invención de la «numeración relativa», atribuída a los sumerios o babilonios y desarrollada por los indios, fué de enorme trascendencia para la civilización. Los sistemas anteriores de numeración estaban basados en un estricto principio aditivo. En el simbolismo romano, p. ej., se escribía

$$\text{CXVIII} = 100 + 10 + 5 + 1 + 1 + 1.$$

Los sistemas egipcio, hebraico y griego eran de un tipo parecido al de los romanos. Un inconveniente de la notación aditiva es que cuanto mayor es el número mayor es también el conjunto de nuevos símbolos necesarios para representarlo. (Claro está que los científicos antiguos no utilizaban las modernas magnitudes astronómicas o atómicas.) Pero el defecto fundamental de los antiguos sistemas, tales como el romano, residía en el hecho de que el cálculo era tan complicado que únicamente los especialistas podían manejar los problemas de cálculo no triviales. Las cosas pasan de modo distinto con el sistema indio de «valor relativo», hoy en uso. Este sistema fué introducido en Europa, en la Edad Media, por comerciantes italianos, quienes lo habían aprendido de los árabes. Tiene la propiedad cómoda de que todos los números, grandes o pequeños, pueden representarse mediante un pequeño conjunto de cifras diferentes (en el sistema decimal, mediante las «cifras árabes» 0, 1, 2, ..., 9). Esto lleva consigo la importante ventaja de la facilidad de los cálculos. Las reglas de cálculo en los sistemas de notación basados en el valor relativo pueden establecerse en forma de tablas de adición y multiplicación para números de una sola cifra y pueden ser aprendidas de memoria y retenidas para siempre. Existen pocos ejemplos de adelantos científicos que hayan afectado tan profundamente y facilitado tanto la vida diaria como el actual sistema de numeración.

3. El cálculo numérico en sistemas distintos del decimal.—El uso del 10 como base de numeración se remonta a los primeros tiempos de la civilización, y es debido indudablemente al hecho de que son diez los dedos con los que se acostumbra contar. Sin embargo, las palabras que sirven para designar algunos números en distintos idiomas parecen reminiscencias del uso de otras bases, especialmente 12 y 20. En inglés y en alemán las palabras para 11 y 12 no están construídas según el principio decimal de combinar 10 con los números de una cifra, como ocurre para los comprendidos entre 13 y 19, sino que son lingüísticamente independientes de la palabra que corresponde a 10. En francés las palabras *vingt* y *quatre-vingt*, para 20 y 80, sugieren la idea

de que para algunas cuestiones se haya usado un sistema de base 20. En danés, la palabra *halvfirsindstyve*, para 70, significa *a medio camino* (a partir de tres veces) de cuatro veces veinte. Los astrónomos de Babilonia utilizaban un sistema que en parte era sexagesimal (base 60); la creencia en este hecho se apoya en la habitual división de la hora y del grado angular en 60 min.

En un sistema distinto del decimal, las reglas de la aritmética son las mismas que en éste; sin embargo, han de usarse otras tablas para la adición y la multiplicación de los números de una cifra. Acostumbrados al sistema decimal y ligados a él por gran número de palabras de nuestro lenguaje, surge al comienzo alguna confusión. Ensayemos un ejemplo de multiplicación en el sistema septimal; antes será conveniente escribir las tablas de sumar y multiplicar que usaremos en la operación:

Adición

	1	2	3	4	5	6
1	2	3	4	5	6	10
2	3	4	5	6	10	11
3	4	5	6	10	11	12
4	5	6	10	11	12	13
5	6	10	11	12	13	14
6	10	11	12	13	14	15

Multiplicación

	1	2	3	4	5	6
1	1	2	3	4	5	6
2	2	4	6	11	13	15
3	3	6	12	15	21	24
4	4	11	15	22	26	33
5	5	13	21	26	34	42
6	6	15	24	33	42	51

Sea multiplicar 265 por 24, símbolos que representan números en el sistema septimal. (En el sistema decimal sería equivalente a multiplicar 145 por 18.) Las reglas de la multiplicación son las mismas que en el sistema decimal. Comenzamos multiplicando 5 por 4, que da 26, como resulta de la tabla de multiplicar.

$$\begin{array}{r} 265 \\ 24 \\ \hline 1456 \\ 563 \\ \hline 10416 \end{array}$$

Escribimos 6 en el lugar de las unidades y «llevamos» 2 al lugar de las septenas. Luego buscamos $4 \cdot 6 = 33$, y $33 + 2 = 35$. Escribimos 5, y procedemos de igual forma hasta que hayamos multiplicado todos los números. Sumando $1456 + 5630$, se obtiene $6 + 0 = 6$ en el lugar de las unidades, $5 + 3 = 11$ en el lugar de las septenas. De nuevo escribimos 1 y conservamos 1 para el lugar de los cuarenta y nueves, con lo que tenemos $1 + 6 + 4 = 14$. El resultado final es: $265 \cdot 24 = 10416$.

Para comprobar este resultado, podemos multiplicar los mismos números en el sistema decimal. 10416 (en el sistema septimal) puede escribirse en el sistema decimal calculando las potencias de 7 hasta la cuarta: $7^2 = 49$, $7^3 = 343$, $7^4 = 2401$. De donde $10416 = 2401 + 4 \cdot 49 + 7 + 6$; este cálculo está hecho en el sistema decimal. Sumando estos números, resulta que 10416 en el sistema septimal es igual a 2610 en el sistema decimal. Luego, multipliquemos 145 por 18 en el sistema decimal; el resultado es 2610, lo que confirma el cálculo anterior.

Ejercicios:

1. Constrúyanse las tablas de adición y multiplicación en el sistema duodecimal y háganse algunos ejemplos en este sistema.
2. Exprésense «30» y «136» en los sistemas de bases, 5, 7, 11 y 12.
3. ¿Qué representan los símbolos 11111 y 21212 en dichos sistemas?
4. Fórmense las tablas de adición y multiplicación en los sistemas de bases 5, 11 y 13.

El sistema de valor relativo de base 2 está caracterizado por ser el de base más pequeña. Las únicas cifras en este *sistema diádico* son 0 y 1; cualquier otro número z vendrá representado por una sucesión de estos dos símbolos. Las tablas de adición y multiplicación se reducen a las reglas $1 + 1 = 10$ y $1 \cdot 1 = 1$. El inconveniente de este sistema es evidente; hacen falta expresiones muy largas para representar números pequeños. Así, 79, que puede expresarse en la forma $1 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2 + 1$, se escribe en el sistema diádico 1001111.

Como ilustración de la sencillez de la multiplicación en el sistema diádico haremos la multiplicación de 7 por 5, números que se representan respectivamente por 111 y 101. Recordando que en este sistema se tiene $1 + 1 = 10$, resulta:

$$\begin{array}{r} 111 \\ \underline{101} \\ 111 \\ \underline{111} \\ 100011 \end{array} = 2^5 + 2 + 1,$$

es decir, 35, como debía ser.

Gottfried Wilhelm Leibniz (1646-1716), una de las mayores inteligencias de su tiempo, fué un apasionado del sistema diádico. A propósito de esto, Laplace escribe: «Leibniz veía en el sistema diádico la imagen de la creación. Consideraba que la unidad representaba a Dios, y el cero, la nada; que el Ser Supremo creaba todos los seres de la nada, del mismo modo que la unidad y el cero expresaban todos los números de su sistema de numeración».

Ejercicio: Considérese la cuestión de representar los enteros en la base a. Para nombrar los enteros en este sistema se necesitan palabras para los números 0, 1, ..., $a - 1$ y para las sucesivas potencias de a: a, a^2, a^3, ... ¿Cuántas palabras distintas son necesarias para designar todos los números de 0 a 1000, para $a = 2$, 3, 4, 5, ..., 15? ¿En qué base son necesarias menos palabras? (Ejemplos: Si $a = 10$, se necesitan diez palabras para los números de una cifra, además de las palabras para 10, 100 y 1000, lo que hace 13 en total. Para $a = 20$, se necesitan veinte para los números de una cifra, más las correspondientes a 20 y 400; en total, 22. Si es $a = 100$, se necesitan 101.)

*II. LA INFINITUD DEL SISTEMA DE NÚMEROS ENTEROS. INDUCCIÓN MATEMÁTICA

1. El principio de inducción matemática.—La sucesión de enteros 1, 2, 3, 4, ..., no tiene fin, puesto que después de cada entero n hay uno siguiente: el $n + 1$. Expresaremos esta propiedad diciendo que la sucesión de enteros contiene *infinitos* enteros. La sucesión de enteros constituye el ejemplo más sencillo y natural del infinito matemático, el cual desempeña un papel dominante en la matemática. A lo largo de este libro necesitaremos manejar colecciones o *conjuntos* que contienen una infinidad de objetos matemáticos; p. ej., el conjunto de todos los puntos de una recta o el conjunto de todos los triángulos de un plano. La sucesión de enteros es el ejemplo más simple de conjunto infinito.

El proceso de ir paso a paso, de n a $n + 1$, que engendra la sucesión infinita de los enteros, forma también la base de uno de los tipos fundamentales de razonamiento matemático: el principio de inducción matemática. La *inducción empírica* de las ciencias naturales procede de una serie particular de observaciones de un cierto fenómeno para establecer una proporción o ley general que debe regir todas las posibilidades del fenómeno. El grado de certeza con que se establece dicha ley depende del número de observaciones particulares y de confirmaciones del fenómeno. Este tipo de razonamiento inductivo es con frecuencia plenamente convincente; la predicción de que mañana el Sol hará su salida por Oriente tiene toda la certeza posible; pero el carácter de esta proposición no es el mismo que el de un teorema probado con razonamientos estrictamente lógicos o matemáticos.

De modo completamente distinto se utiliza la *inducción matemática* para establecer la certeza de un teorema matemático en una sucesión infinita de casos: el primero, el segundo, el tercero, y así sucesivamente, sin excepción. Designemos con A una proposición que se refiera a un entero arbitrario n; p. ej., A puede ser la proposición: «La suma de los ángulos de un polígono convexo de $n + 2$ lados es n veces 180°.» A' puede consistir en la afirmación: «Trazando n rectas en

un plano no se puede dividir éste en más de 2^n partes.» Para probar uno de estos teoremas para *cualquier* entero n no es suficiente probarlo para los 10 ó 100, ni aun para los 1000 primeros valores de n. Este modo de proceder correspondería precisamente a la inducción empírica. En su lugar, debemos usar un método estrictamente matemático y no un razonamiento empírico, cuyo carácter indicaremos en lo que sigue, al probar los ejemplos especiales A y A'. En el caso A sabemos que para $n = 1$ el polígono es un triángulo, y por geometría elemental se sabe que la suma de sus ángulos es $1 \cdot 180^o$. Para un cuadrilátero, $n = 2$, se traza una diagonal que dividirá al cuadrilátero en dos triángulos. Entonces se ve, de modo inmediato, que la suma de los ángulos del cuadrilátero es igual a la suma de los ángulos de los dos triángulos, lo que da $180^o + 180^o = 2 \cdot 180^o$. Procediendo análogamente para el pentágono, $n = 3$, se descompone en un triángulo y un cuadrilátero. Puesto que la suma de los ángulos del último es $2 \cdot 180^o$, como acabamos de probar, y siendo la suma de los ángulos del triángulo 180^o, obtenemos para el pentágono $3 \cdot 180^o$. Ahora bien: resulta claro que podemos proceder indefinidamente en la misma forma, probando el teorema para $n = 4$; luego, para $n = 5$, y así sucesivamente. Cada proposición se deduce de la precedente en la misma forma, de modo que el teorema general puede ser establecido para todo n.

Análogamente podemos probar A'; para $n = 1$ es evidente, ya que una recta divide al plano en dos partes. Añadamos una segunda recta. Cada una de las partes anteriores quedará dividida en dos nuevas partes, salvo que la nueva recta sea paralela a la primera. En ambos casos, para $n = 2$ no resultan más de $4 = 2^2$ partes. Añadamos una tercera recta; cualquiera de las regiones anteriores quedará, o bien dividida en dos partes o sin dividir. Así, la suma de partes no podrá ser mayor que $2 \cdot 2^2 = 2^3$. Sabiendo que esto es cierto, podemos probar el caso siguiente en la misma forma, y así indefinidamente.

La idea esencial de los argumentos precedentes consiste en establecer un teorema general A para todos los valores de n, probando sucesivamente una sucesión de casos especiales A_1, A_2, ... La posibilidad de hacerlo depende de dos hechos: *a)* Existe un método general para probar que si cualquier proposición A_r es cierta, la proposición siguiente, A_{r+1}, será *también* cierta. *b)* Se *sabe* que la primera proposición A es cierta. Que estas dos condiciones son suficientes para establecer la verdad de *todas* las proposiciones A_1, A_2, A_3, ..., es un principio lógico que resulta tan fundamental para las matemáticas como lo son las reglas clásicas de la lógica aristotélica. Por ello vamos a enunciarlo explícitamente como sigue:

Supongamos que queremos establecer una sucesión infinita de proposiciones matemáticas

$$A_1, A_2, A_3, \ldots$$

que juntas constituyen una proposición general A. *Supongamos:* a) *que por un razonamiento matemático se prueba que, si* r *es un entero cualquiera, de la verdad de la proposición* A_r *se sigue la verdad de la* A_{r+1}, *y* b), *se sabe que la proposición* A_1 *es cierta. Entonces todas las proposiciones de la sucesión son ciertas y queda probada* A.

No debe haber duda en aceptar esto, del mismo modo que no la tenemos para aceptar las reglas elementales de la lógica ordinaria, como un principio del razonamiento matemático, ya que se puede establecer la verdad de cualquiera de las proposiciones A_n partiendo de la aserción *b)* de que A_1 es cierta, y procediendo por uso repetido de la aserción *a)*, para establecer sucesivamente la verdad de A_2, A_3, A_4, y así hasta llegar a la proposición A_n. El principio de inducción matemática se basa en el hecho de que después de cada entero r hay un siguiente $r + 1$, y que todo entero n puede ser alcanzado mediante un número finito de pasos, a partir del 1.

Con frecuencia, el principio de inducción matemática se aplica sin mencionarlo explícitamente, o viene indicado simplemente por un «etc.» o un «y así sucesivamente». Así sucede, en particular, en la enseñanza elemental. Pero el uso explícito del razonamiento inductivo es indispensable en demostraciones más sutiles. Damos a continuación algunos ejemplos de carácter sencillo, pero no trivial.

2. Progresiones aritméticas.—*Para todo valor de* n, *la suma* $1 + 2 + 3 \ldots + n$ *de los* n *primeros enteros es igual a* $\frac{n(n+1)}{2}$. Para probar este teorema por inducción matemática debemos demostrar que para cualquier n la proposición A_n:

$$1 + 2 + 3 + \ldots + n = \frac{n(n+1)}{2} \qquad [1]$$

es cierta. *a)* Observemos que si r es un entero y si se sabe que la proposición A_r es cierta; es decir, si sabemos que

$$1 + 2 + 3 + \ldots + r = \frac{r(r+1)}{2},$$

sumándole el número $(r + 1)$ a los dos miembros de esta igualdad obtenemos la ecuación

$$1 + 2 + 3 + \ldots + r + (r + 1) = \frac{r(r+1)}{2} + (r + 1) = \frac{(r+1)(r+2)}{2}$$

que es precisamente la proposición A_{r+1}. *b)* La proposición A_1 es evidente, ya que $1 = \frac{1 \cdot 2}{2}$. De donde, por el principio de inducción matemática, la proposición A_n es cierta para todo n, como se quería demostrar.

Corrientemente se suele probar escribiendo la suma $1 + 2 + 3 + + \dots + n$ de dos maneras:

$$S_n = 1 + 2 + \dots + (n - 1) + n$$

y

$$S_n = n + (n - 1) + \dots + 2 + 1.$$

Al sumar, se observa que cada par de números de la misma columna da como suma $n + 1$, y puesto que hay n columnas en total, se sigue

$$2S_n = n(n + 1),$$

lo que prueba el resultado indicado.

De [1] se puede deducir de modo inmediato la fórmula de la suma de los $(n + 1)$ primeros términos de cualquier *progresión aritmética*,

$$P_n = a + (a + d) + (a + 2d) + \dots + (a + nd) = \frac{(n + 1)(2a + nd)}{2} \qquad [2]$$

Puesto que

$$P_n = (n + 1)a + (1 + 2 + \dots + n)d = (n + 1)a + \frac{n(n + 1)d}{2} =$$

$$= \frac{2(n + 1)a + n(n + 1)d}{2} = \frac{(n + 1)(2a + nd)}{2}$$

Para el caso $a = 0$ y $d = 1$, esta fórmula es la misma [1]:

3. Progresiones geométricas.—Se pueden estudiar las progresiones geométricas de modo análogo al precedente. Probaremos que para todo valor de n se tiene

$$G_n = a + aq + aq^2 + \dots + aq^n = a\frac{1 - q^{n+1}}{1 - q} \qquad [3]$$

(Suponemos $q \neq 1$, ya que de otro modo el último miembro de [3] no tendría significado.)

La proposición es cierta para $n = 1$, ya que entonces

$$G_1 = a + aq = \frac{a(1 - q^2)}{1 - q} = \frac{a(1 + q)(1 - q)}{(1 - q)} = a(1 + q).$$

Y *si* suponemos que se tiene

$$G_r = a + aq + \dots + aq^r = a\frac{1 - q^{r+1}}{1 - q},$$

resulta como consecuencia

$$G_{r+1} = (a + aq + \dots + aq^r) + aq^{r+1} = G_r + aq^{r+1} = a\frac{1 - q^{r+1}}{1 - q} + aq^{r+1} =$$

$$= a\frac{(1 - q^{r+1}) + q^{r+1}(1 - q)}{1 - q} = a\frac{1 - q^{r+1} + q^{r+1} - q^{r+2}}{1 - q} = a\frac{1 - q^{r+2}}{1 - q}$$

Pero esto es precisamente la proposición [3] para el caso $n = r + 1$, lo que completa la demostración.

En los libros elementales, la prueba habitual procede como sigue. Pongamos

$$G_n = a + aq + \dots + aq^n,$$

y multipliquemos los dos miembros por q. Se obtiene

$$qG_n = aq + aq^2 + \dots + aq^{n+1}.$$

Restando entonces los miembros correspondientes de las dos igualdades, resulta

$$G_n - qG_n = a - aq^{n+1},$$
$$(1 - q)G_n = a(1 - q^{n+1}),$$
$$G_n = a\frac{1 - q^{n+1}}{1 - q}$$

4. Suma de los «n» primeros cuadrados.—Otra interesante aplicación del principio de inducción se refiere a la suma de los n primeros cuadrados. Mediante ensayos directos se encuentra, al menos para valores pequeños de n,

$$1^2 + 2^2 + 3^2 + \dots + n^2 = \frac{n(n + 1)(2n + 1)}{6} \qquad [4]$$

y se *supone* que esta fórmula pueda ser válida para *todos los enteros n*. Para *probarlo*, haremos uso de nuevo del principio de inducción. Comenzaremos por observar que si la proposición A_n, que en este caso coincide con la ecuación [4], es válida para el caso $n = r$, de modo que se tenga

$$1^2 + 2^2 + 3^2 + \dots + r^2 = \frac{r(r + 1)(2r + 1)}{6},$$

sumando $(r + 1)^2$ a los dos miembros de esta igualdad se obtiene

$$1^2 + 2^2 + 3^2 + \dots + r^2 + (r + 1)^2 = \frac{r(r + 1)(2r + 1)}{6} + (r + 1)^2 =$$

$$= \frac{r(r + 1)(2r + 1) + 6(r + 1)^2}{6} = \frac{(r + 1)[r(2r + 1) + 6(r + 1)]}{6} =$$

$$= \frac{(r + 1)(2r^2 + 7r + 6)}{6} = \frac{(r + 1)(r + 2)(2r + 3)}{6},$$

que es precisamente la proposición A_{r+1} en el caso presente, puesto que se obtiene sustituyendo n por $r + 1$ en [4]. Para completar la demostración necesitamos únicamente observar que la proposición A_1, en este caso la ecuación

$$1^2 = \frac{1(1 + 1)(2 + 1)}{6},$$

es evidente. Por tanto, la ecuación [4] es válida para cualquier n.

Fórmulas análogas pueden hallarse para potencias superiores de los enteros, $1^k + 2^k + 3^k + ... + n^k$, donde k es un entero positivo cualquiera. Como ejercicio, el lector puede probar, mediante inducción matemática, que

$$1^3 + 2^3 + 3^3 + ... + n^3 = \left[\frac{n(n + 1)}{2}\right]^2 \qquad [5]$$

Debe observarse que, si bien el principio de inducción matemática es suficiente para *probar* la fórmula [5] una vez que se conoce ésta, la demostración no da indicación alguna sobre el modo en que dicha fórmula puede encontrarse; es decir, sobre el por qué debe suponerse la expresión $[n(n + 1)/2]^2$ como resultado para la suma de los n primeros cubos, en vez de la $[n(n + 1)/3]^2$ o $(19n^2 - 41n + 24)/2$ o cualquiera de las infinitas expresiones análogas que pudieran ser consideradas. El hecho de que la demostración de un teorema consista en la aplicación de ciertas reglas sencillas de lógica no disminuye el valor del elemento creador en matemáticas, el cual desempeña su papel en la elección de las posibilidades que deban ser tenidas en cuenta. La cuestión del origen de la *hipótesis* [5] pertenece a un dominio en el cual no pueden ser dadas reglas generales; ensayos, analogías e intuición constructiva tienen en esto papel importante. Pero una vez formulada correctamente la hipótesis, el principio de inducción matemática es con frecuencia suficiente para dar la demostración. En tanto que una demostración no proporcione una indicación para el acto del descubrimiento, debe llamarse más propiamente una *comprobación*.

***5. Una desigualdad importante.**—En otro capítulo haremos uso de la desigualdad

$$(1 + p)^n \geqslant 1 + np, \qquad [6]$$

que es válida para todo número $p > -1$ y todo entero positivo n. (Con objeto de dar mayor generalidad, anticipamos aquí el uso de números negativos no enteros admitiendo como p cualquier número mayor que -1. La demostración en el caso general es exactamente

la misma que en el caso en que p es un entero positivo.) Usaremos de nuevo la inducción matemática.

a) Si es cierto que $(1 + p)^r \geqslant 1 + rp$, multiplicando los dos miembros de esta desigualdad por el número positivo $1 + p$ se obtiene

$$(1 + p)^{r+1} > 1 + rp + p + rp^2.$$

Si prescindimos del término positivo rp^2, se tendrá

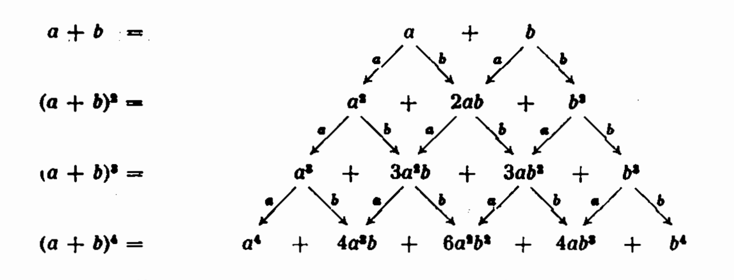
$$(1 + p)^{r+1} > 1 + (r + 1)p,$$

lo que prueba que la desigualdad [6] vale también para el entero siguiente $r + 1$. *b)* Evidentemente se tiene: $(1 + p)^1 \geqslant 1 + p$. Esto completa la demostración de que [6] es cierta para todo n. La restricción de que $p > -1$ es esencial. Si es $p < -1$, $1 + p$ es negativo y el razonamiento empleado para *a)* deja de ser válido, puesto que si se multiplican los dos miembros de una desigualdad por un número negativo, la desigualdad cambia de sentido. (P. ej., si multiplicamos los dos miembros de la desigualdad $3 > 2$ por -1, obtendríamos $-3 < -2$.)

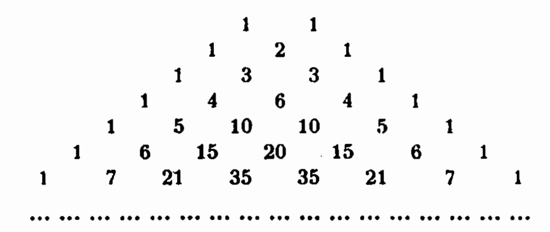
***6. El binomio de Newton.**—Con frecuencia es importante tener una expresión explícita para la n-ésima potencia de un binomio, $(a + b)^n$. Mediante cálculo explícito se obtiene:

para $n = 1$, $(a + b)^1 = a + b$.
para $n = 2$, $(a + b)^2 = (a + b)(a + b) = a(a + b) + b(a + b) = a^2 + 2ab + b^2$,
para $n = 3$, $(a + b)^3 = (a + b)(a + b)^2 = a(a^2 + 2ab + b^2) + b(a^2 + 2ab + b^2) = a^3 + 3a^2b + 3ab^2 + b^3$,

y así sucesivamente. ¿Cuál es la ley general de formación, implícita en la frase «y así sucesivamente»? Examinemos el proceso mediante el cual calculamos $(a + b)^2$. Puesto que es $(a + b)^2 = (a + b)(a + b)$, se obtiene el desarrollo de $(a + b)^2$ multiplicando primero cada término de la expresión $a + b$ por a y luego por b y sumando después. El mismo procedimiento se utilizó para calcular $(a + b)^3 = (a + b)(a + b)^2$. Se puede continuar del mismo modo para calcular $(a + b)^4$, $(a + b)^5$, y así indefinidamente. La expresión de $(a + b)^n$ se obtendría multiplicando cada uno de los términos de la expresión de $(a + b)^{n-1}$, calculada previamente, primero por a y luego por b, y sumando. Esto conduce al diagrama



el cual da inmediatamente la regla general para formar los coeficientes del desarrollo de $(a + b)^n$. Para ello construímos un esquema triangular de números, partiendo de los coeficientes 1, 1 de $a + b$, y de tal modo que cada número del triángulo es la suma de los dos números inmediatos a él en la fila precedente. Esta disposición de los números es conocida con el nombre de *triángulo de Pascal.*



La fila n*-ésima de este esquema da los coeficientes del desarrollo de* $(a + b)^n$ *según las potencias decrecientes de* a *y las crecientes de* b; así

$$(a + b)^7 = a^7 + 7a^6b + 21a^5b^2 + 35a^4b^3 + 35a^3b^4 + 21a^2b^5 + 7ab^6 + b^7.$$

Usando una notación concisa, mediante índices y subíndices, se pueden representar los números de la n-ésima fila del triángulo de Pascal por

$$C_0^n = 1,\ C_1^n,\ C_2^n,\ C_3^n,\ \ldots,\ C_{n-1}^n,\ C_n^n = 1.$$

Entonces, la fórmula general para $(a + b)^n$ puede escribirse

$$(a + b)^n = a^n + C_1^n a^{n-1}b + C_2^n a^{n-2}b^2 + \ldots + C_{n-1}^n ab^{n-1} + b^n. \qquad [7]$$

De acuerdo con la ley de formación del triángulo de Pascal, se tiene

$$C_i^n = C_{i-1}^{n-1} + C_i^{n-1}. \qquad [8]$$

Como ejercicio, el lector puede utilizar esta relación, junto con el

hecho de que $C_0^1 = C_1^1 = 1$, para probar, mediante inducción matemática, que

$$C_i^n = \frac{n(n-1)(n-2)\dots(n-i+1)}{1\cdot 2\cdot 3\dots i} = \frac{n!}{i!(n-i)!} \qquad [9]$$

(Para todo entero positivo n, el símbolo $n!$—léase «factorial de n»—designa el producto de los n primeros enteros: $n! = 1\cdot 2\cdot 3\cdots n$. Y resulta conveniente definir también $0!$ por la igualdad $0! = 1$; de este modo [9] es válido para $i = 0$ e $i = n$.) Esta fórmula explícita para los coeficientes del desarrollo binómico se conoce con el nombre de *teorema del binomio*. (Véase también Cap. VIII, Suplemento, III, 1.)

Ejercicios: Demuéstrese por inducción matemática:

1. $\frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \dots + \frac{1}{n(n+1)} = \frac{n}{n+1}$

2. $\frac{1}{2} + \frac{2}{2^2} + \frac{3}{2^3} + \dots + \frac{n}{2^n} = 2 - \frac{n+2}{2^n}$

*3. $1 + 2q + 3q^2 + \dots + nq^{n-1} = \frac{1-(n+1)q^n + nq^{n+1}}{(1-q)^2}$

*4. $(1+q)(1+q^2)(1+q^4)\dots(1+q^{2^n}) = \frac{1-q^{2^{n+1}}}{1-q}$

Hállese la suma de las siguientes progresiones geométricas:

5. $\frac{1}{1+x^2} + \frac{1}{(1+x^2)^2} + \dots + \frac{1}{(1+x^2)^n}$

6. $1 + \frac{x}{1+x^2} + \frac{x^2}{(1+x^2)^2} + \dots + \frac{x^n}{(1+x^2)^n}$

7. $\frac{x^2-y^2}{x^2+y^2} + \left(\frac{x^2-y^2}{x^2+y^2}\right)^2 + \dots + \left(\frac{x^2-y^2}{x^2+y^2}\right)^n$

Utilizando las fórmulas [4] y [5], pruébese:

*8. $1^2 + 3^2 + \dots + (2n+1)^2 = \frac{(n+1)(2n+1)(2n+3)}{3}$

*9. $1^3 + 3^3 + \dots + (2n+1)^3 = (n+1)^2(2n^2+4n+1)$.

10. Demuéstrense los mismos resultados directamente por inducción matemática.

***7. Algunas observaciones a propósito de la inducción matemática.**—El principio de inducción matemática puede modificarse un poco y darle la siguiente forma:

«Si una sucesión de proposiciones $A_s, A_{s+1}, A_{s+2}, \ldots$, donde s es un entero positivo, es tal que:

a) para todo $r \geqslant s$, de la verdad de A_r se sigue la de A_{r+1}, y

b) se sabe que A_s es cierta,

entonces todas las proposiciones $A_s, A_{s+1}, A_{s+2}, \ldots$ son ciertas; es decir, A_n es cierta para todo $n \geqslant s$.» Se aplica aquí el mismo razonamiento utilizado para establecer la validez del principio ordinario de inducción matemática, con la única variante de sustituir la sucesión 1, 2, 3, ... por la sucesión análoga, $s, s+1, s+2, \ldots$ Usando el principio en esta forma, se puede precisar más la desigualdad de la página 22, eliminando la posibilidad del signo «=». Se puede probar que: *Para todo* $p \neq 0$ y > -1 *y para todo entero* $n \geqslant 2$,

$$(1+p)^n > 1 + np. \qquad [10]$$

Dejamos la demostración al cuidado del lector.

Íntimamente ligado con el principio de inducción matemática está el llamado «principio del menor entero», el cual establece *que todo conjunto* C, *no vacío, de números enteros positivos contiene un entero menor que todos los demás.* Un conjunto vacío es aquel que no contiene elementos; p. ej., el conjunto de las circunferencias rectilíneas o el conjunto de los enteros n tales que $n > n$. Por razones obvias excluímos tales conjuntos de nuestro principio. El conjunto C puede ser finito, tal como el conjunto 1, 2, 3, 4, 5, ó infinito, como el conjunto de todos los números pares 2, 4, 6, 8, 10, ... Cualquier conjunto C, no vacío, debe contener cuando menos un entero, p. ej., el n, y entonces el más pequeño de los enteros 1, 2, 3, ..., n, que pertenezca a C será el menor de los enteros contenidos en C.

Para comprender bien el significado de este principio se debe observar que deja de ser cierto si se aplica a cualquier conjunto C de números que no sean enteros; p. ej., el conjunto de fracciones positivas $1, {}^1/_2, {}^1/_3, \ldots$ no contiene una fracción menor que todas las demás.

Desde el punto de vista de la lógica es interesante observar que el principio del menor entero puede ser utilizado para *demostrar* el principio de inducción matemática como un teorema. Con tal objeto, consideremos cualquier sucesión de proposiciones $A_1, A_2, A_3, \ldots$ tales que:

a) Para todo entero positivo r, la verdad de A_{r+1} se sigue de la de A_r.

b) Se sabe que A_1 es cierta.

Probaremos que la hipótesis de que cualquier A sea falsa es absurda. Pues si alguna de las A fuese falsa, el conjunto C de todos los enteros positivos n para los cuales fuera falsa A_n sería un conjunto no vacío. Por el principio del menor entero, C contendría un entero p más pequeño que todos los demás, y p debería ser > 1 a causa de *b)*. Por tanto, se tendría que A_p sería falsa mientras que A_{p-1} sería cierta; pero esto contradice *a)*.

Una vez más insistimos en que el principio de inducción matemática es completamente diferente de la inducción empírica de las ciencias naturales. La confirmación de una ley general en cualquier número finito de casos, por grande que sea dicho número, no suministra una demostración de la ley en sentido matemático riguroso, y esto aunque no sea conocida excepción alguna. Tal ley quedará únicamente como una *hipótesis* plausible, sujeta siempre a modificaciones por los

resultados de ulteriores experiencias. En matemáticas, una ley o un teorema quedan probados únicamente cuando se demuestra que son una consecuencia lógicamente necesaria de ciertos supuestos admitidos como válidos. Existen varios ejemplos de proposiciones matemáticas que hàn sido comprobadas en todos los numerosos casos particulares considerados, pero que hasta la fecha no han sido demostradas en general (para un ejemplo, véase pág. 38). Se puede *sospechar* que un teorema es cierto en general si resulta cierto en un número de ejemplos; entonces cabe intentar *probarlo* mediante la inducción matemática. Si el intento tiene éxito, el teorema queda demostrado; si se fracasa, el teorema puede ser cierto o falso y algún día podrá, posiblemente por otros métodos, ser probado o rechazado.

Al usar el principio de inducción matemática se debe estar seguro de que las condiciones *a)* y *b)* están efectivamente satisfechas. Descuidar esta precaución puede conducir a absurdos como el que sigue (invitamos al lector para que descubra el error en el razonamiento). «Probaremos» que *dos enteros positivos cualesquiera son iguales;* p. ej., que 5 = 10.

Comenzamos con una definición: Si a y b son dos enteros positivos desiguales, definimos como máx. (a, b) aquel de los dos que sea mayor: si es $a = b$ pondremos máx. $(a, b) = a = b$. Así máx. (3, 5) = máx. (5, 3) = 5, mientras que máx. (4, 4) = 4. Sea ahora A_n la proposición: «Si a y b son dos enteros positivos cualesquiera, tales que máx. $(a, b) = n$, se tiene $a = b$.»

a) Supongamos que A_r es cierta. Sean a y b dos enteros positivos cualesquiera, tales que máx. $(a, b) = r + 1$. Consideremos los dos enteros

$$\alpha = a - 1$$

$$\beta = b - 1;$$

se tendrá entonces: máx. $(\alpha, \beta) = r$. De aquí resulta $\alpha = \beta$, puesto que suponíamos que A_r era cierta. De donde se sigue $a = b$; en consecuencia, A_{r+1} es cierta.

b) A_1 es evidentemente cierta, pues si es máx. $(a, b) = 1$, se tendrá, ya que suponemos a y b enteros positivos, que estos dos números deben ser iguales a 1. Por consiguiente, en virtud de la inducción matemática, A_n es cierta para todo n.

Ahora, si a y b son dos enteros positivos cualesquiera, designemos máx. (a, b) por r. Puesto que hemos probado que A_n es cierta para todo n, se tendrá en particular que A_r es cierta. En consecuencia $a = b$.

SUPLEMENTO AL CAPÍTULO PRIMERO

TEORÍA DE NÚMEROS

Introducción.—Los enteros fueron poco a poco perdiendo su relación con supersticiones y misticismos, pero su interés para los matemáticos no disminuyó nunca. Euclides (hacia el año 300 a. de J.C.), cuya fama está ligada a la parte de sus *Elementos* que forma la base de la geometría elemental, parece haber obtenido resultados originales en la teoría de números, mientras que su geometría es, en su mayor parte, una compilación de resultados anteriores. Diofanto de Alejandría (hacia el año 275 d. de J.C.), uno de los primeros algebristas, dejó huella importante en la teoría de números. Pierre de Fermat (1601-1665), jurista en Toulouse y uno de los más grandes matemáticos de su tiempo, inició las investigaciones modernas en este campo. Euler (1707-1783), el más prolífico de los matemáticos, obtuvo gran cantidad de resultados en la teoría de números entre sus investigaciones matemáticas. Otros nombres preeminentes en los anales de la matemática—Legendre, Dirichlet, Riemann—deben añadirse a la lista anterior. Gauss (1777-1855), el más célebre matemático de los tiempos modernos, dedicado asimismo a distintas ramas de las matemáticas, resumió su opinión sobre la teoría de números en la frase: «La matemática es la reina de las ciencias, y la teoría de números es la reina de las matemáticas.»

I. LOS NÚMEROS PRIMOS

1. Hechos fundamentales.—La mayor parte de las proposiciones de la teoría de números, como en general de la matemática, no se refieren a un objeto particular—el número 5 ó el número 32—, sino a una clase completa de objetos que tienen alguna propiedad común; p. ej., la clase de todos los números pares:

$$2, 4. 6, 8, \ldots,$$

o la clase de los enteros divisibles por 3:

$$3, 6, 9, 12, \ldots,$$

o bien la clase de los cuadrados de los números enteros:

$$1, 4, 9, 16, \ldots,$$

y así sucesivamente.

De fundamental importancia en teoría de números es la clase de los números *primos*. Casi todos los enteros se pueden descomponer en producto de factores más pequeños: $10 = 2 \cdot 5$, $111 = 3 \cdot 37$, $144 = 3 \cdot 3 \cdot 2 \cdot 2 \cdot 2 \cdot 2$, etc. Los números que no pueden descomponerse de este modo se llaman números primos o simplemente primos. Con más precisión, *un número primo es un entero* p, *mayor que uno, que no admite más factores que él mismo y la unidad.* (Se dice que un *entero a* es *factor* o *divisor* de otro *entero b* si existe otro *entero c* tal que $b = ac$.) Los números 2, 3, 5, 7, 11, 13, 17, ... son primos, mientras que 12, p. ej., no es primo, ya que se tiene $12 = 3 \cdot 4$. La importancia de la clase de los números primos se debe al hecho de que cualquier entero puede expresarse como *producto de números primos*: si un número no es primo, se puede descomponer sucesivamente hasta que todos sus factores sean primos; p. ej., $360 = 3 \cdot 120 = 3 \cdot 30 \cdot 4 = 3 \cdot 3 \cdot 10 \cdot 2 \cdot 2 = 3 \cdot 3 \cdot 5 \cdot 2 \cdot 2 \cdot 2 = 2^3 \cdot 3^2 \cdot 5$. Un entero (distinto de 0 y 1) que no sea primo se llama *compuesto*.

Una de las primeras cuestiones que se presentan en la teoría de los números primos es la de saber si hay solamente un número finito de estos números o si, por el contrario, la clase de los números primos tiene infinitos elementos distintos, como ocurre con la clase de los números enteros de la cual forma parte. La respuesta es: *Existen infinitos números primos.*

La prueba de la infinitud de la clase de los números primos dada por Euclides ha quedado como modelo de razonamiento matemático. Procede por «reducción al absurdo», y parte de la hipótesis de que el teorema es falso. Esto significa que existiría únicamente un número finito de números primos, quizá muchos (p. ej., un millón) o, dicho de un modo general, n. Usando subíndices, podríamos indicar dichos números primos por $p_1, p_2, \ldots, p_n$. Cualquier otro entero sería compuesto y debería ser divisible por uno al menos de los $p_1, p_2, \ldots, p_n$. Vamos a ver que dicha hipótesis nos lleva a una contradicción; para ello, construyamos un número A que será distinto de los primos $p_1, p_2, \ldots, p_n$ por ser mayor que ellos y que, sin embargo, no será divisible por ninguno de los p. Tal número es el

$$A = p_1 p_2 \ldots p_n + 1,$$

es decir, se obtiene añadiendo 1 al producto de todos los números primos que suponíamos existentes. A es distinto de los p y, por tanto, debe ser compuesto. Pero dividiendo A por p_1, o por p_2, etc., se obtiene siempre de resto 1; en consecuencia, A no admite ningún p como divisor. Puesto que nuestra hipótesis inicial de que había sólo

un número finito de números primos nos conduce a una contradicción, dicha hipótesis es absurda y, por tanto, la contraria debe ser cierta; esto prueba el teorema.

Aunque la demostración es indirecta, puede modificarse fácilmente para dar un método constructivo, al menos teóricamente, de una sucesión infinita de números primos. Partiendo de un número primo cualquiera, p. ej., $p_1 = 2$, supongamos que hemos encontrado n primos $p_1, p_2, p_3, \ldots, p_n$; observemos entonces que el número $p_1p_2 \ldots p_n + 1$, o bien es primo o contiene factores primos que han de ser distintos de los n hallados previamente. Puesto que estos factores pueden hallarse por ensayos directos, estamos seguros de que, en todo caso, hay al menos un nuevo factor primo p_{n+1}; procediendo de este modo se ve que la sucesión de los números primos construíbles no tiene fin.

Ejercicio: Llévese a efecto dicha construcción a partir de $p_1 = 2$, $p_2 = 3$ y obténganse 5 números primos más.

Si un número ha sido expresado como producto de números primos, podemos disponer dichos factores primos en un orden cualquiera. La experiencia demostraría que, salvo la arbitrariedad en la ordenación, la descomposición de un número N en factores primos es única: *Todo entero* N, *mayor que* 1, *puede descomponerse en producto de números primos, y solamente de una forma.* Esta proposición parece a simple vista tan evidente que un profano podría inclinarse a admitirla sin prueba. Sin embargo, no es una trivialidad y la demostración, aunque elemental, requiere algunos razonamientos sutiles. La demostración clásica, dada por Euclides, de este «teorema fundamental de la aritmética» está basada en un método o «algoritmo» para el cálculo del máximo común divisor de dos números. Este método será considerado en la página 51. En vez de dicha demostración, daremos aquí otra de cosecha más reciente; más breve, pero quizá más artificiosa que la de Euclides. Será un ejemplo típico de demostración indirecta. Supondremos la existencia de un entero susceptible de dos descomposiciones esencialmente diferentes, y de esta hipótesis resultará una contradicción. Esta contradicción demostrará que la hipótesis de que existe un entero con dos descomposiciones esencialmente diferentes en factores primos es absurda, y, como consecuencia, resultará que la descomposición en factores primos de un entero cualquiera es única.

*Si existe un entero positivo capaz de descomponerse en dos productos esencialmente diferentes de primos, habrá uno *menor* que todos los demás para el que se verifique tal propiedad (véase pág. 26),

$$m = p_1p_2 \ldots p_r = q_1q_2 \ldots q_s, \qquad [1]$$

donde los p y los q son primos. Ordenando de nuevo, si es preciso, los p y los q, podemos suponer que se tiene

$$p_1 < p_2 < \cdots < p_r, \qquad q_1 < q_2 < \cdots < q_s.$$

Ahora bien: p_1 no puede ser igual a q_1, ya que en caso contrario, dividiendo los dos últimos miembros de [1] por $p_1 = q_1$, se obtendrían dos descomposiciones esencialmente diferentes para un entero menor que m, en contradicción con la elección hecha de m como el entero *más pequeño* para el cual ese hecho es posible. Por tanto, o bien es $p_1 < q_1$ o se tiene $q_1 < p_1$. Supongamos $p_1 < q_1$. (Si fuera $q_1 < p_1$, bastaría cambiar las letras p y q en lo que sigue.) Formemos el entero

$$m' = m - (p_1 q_2 q_3 \cdots q_s). \qquad [2]$$

Sustituyendo m por las dos expresiones dadas por [1], podríamos escribir el entero m' en una de las dos formas

$$m' = (p_1 p_2 \cdots p_r) - (p_1 q_2 \cdots q_s) = p_1(p_2 p_3 \cdots p_r - q_2 q_3 \cdots q_s) \qquad [3]$$

$$m' = (q_1 q_2 \cdots q_s) - (p_1 q_2 \cdots q_s) = (q_1 - p_1)(q_2 q_3 \cdots q_s) \qquad [4]$$

Puesto que es $p_1 < q_1$, de [4] se sigue que m' es un entero positivo, mientras que de [2] resulta que m' es menor que m. De donde se deduce que, salvo el orden de los factores, la descomposición de m' debe ser *única*. Pero de [3] resulta que p_1 es un factor de m'; por tanto, de [4] se concluye que p_1 debe aparecer como factor o del $(q_1 - p_1)$ o de $(q_2 q_3 \ldots q_s)$. (Esto resulta de suponer la descomposición única para m'; véase el razonamiento del párrafo siguiente.) La última hipótesis es imposible, porque p_1 es menor que todas las q. En consecuencia, p_1 debe ser un factor de $q_1 - p_1$, de modo que existirá un h tal que

$$q_1 - p_1 = p_1 \cdot h \qquad \text{o} \qquad q_1 = p_1(h + 1).$$

Pero esto expresa que p_1 es un factor de q_1, contrariamente al hecho de ser q_1 número primo. Esta contradicción prueba que nuestra hipótesis inicial era absurda y, por tanto, completa la demostración del teorema fundamental de la aritmética.

Un corolario importante de este teorema fundamental es el siguiente: *Si un número primo* p *es un divisor del producto* ab, p *debe ser factor de* a *o de* b. Pues si p no fuera divisor de a ni de b, el producto de factores primos que da la descomposición de ab *no contendría* p. Por otra parte, puesto que se supone que p es un divisor de ab, existirá un entero t tal que

$$ab = pt.$$

En consecuencia, el producto de p por una descomposición en factores primos de t daría una descomposición de ab en factores primos, que *contendría* p, lo que estaría en contradicción con el hecho de que la descomposición en factores primos es única.

Ejemplos.—Si se ha comprobado de una parte que 13 es un divisor de 2652, y de otra que es $2652 = 6 \cdot 442$, se puede concluir que 13 es un divisor de 442. Por el contrario, 6 es un factor de 240, y $240 = 15 \cdot 16$; sin embargo, 6 no es factor de 15 ni de 16. Esto último prueba que la hipótesis de que p es *primo* es esencial para el corolario.

Ejercicio: Para hallar todos los divisores de un número cualquiera a es suficiente descomponer a en un producto

$$a = p_1^{\alpha_1} \cdot p_2^{\alpha_2} \cdots p_r^{\alpha_r},$$

donde las p son los distintos factores primos, cada uno de ellos elevado a una cierta potencia. *Todos* los divisores de a son los números

$$b = p_1^{\beta_1} \cdot p_2^{\beta_2} \cdots p_r^{\beta_r},$$

donde las β son enteros cualesquiera que satisfacen las desigualdades

$$0 < \beta_1 < \alpha_1, \quad 0 < \beta_2 < \alpha_2, \ldots, \quad 0 < \beta_r < \alpha_r.$$

Demuéstrese esta proposición. Como consecuencia, pruébese que el número de divisores distintos de a (incluidos los divisores a y 1) viene dado por el producto

$$(\alpha_1 + 1)(\alpha_2 + 1) \cdots (\alpha_r + 1).$$

Por ejemplo,

$$144 = 2^4 \cdot 3^2$$

tiene $5 \cdot 3$ divisores, que son: 1, 2, 4, 8, 16, 3, 6, 12, 24, 48, 9, 18, 36, 72 y 144.

2. Distribución de los números primos.—Puede construirse una lista de todos los números primos menores que un entero N dado escribiendo primero ordenadamente todos los enteros menores que N, tachando después todos los que sean múltiplos de 2; luego, todos los restantes que sean múltiplos de 3, y así sucesivamente, hasta que hayan sido eliminados todos los números compuestos. Este proceso, conocido como la «criba de Eratóstenes», conserva en el tamiz los números primos menores que N. Tablas completas de números primos hasta 10 000 000 han sido calculadas mediante perfeccionamientos del método anterior; gracias a ellas se dispone de una masa imponente de datos empíricos referentes a la distribución y otras propiedades de los números primos. Sobre la base de estas tablas se han adelantado conjeturas plausibles (del mismo modo que si la teoría de números fuese una ciencia experimental), muchas de las cuales resultan de demostración muy difícil.

a) *Fórmulas que dan números primos.*—Se han hecho muchos intentos para obtener fórmulas aritméticas simples que dieran única-

mente números primos, aunque no se obtuvieran todos. Fermat hizo la famosa conjetura (no en forma de afirmación definitiva) de que todos los números de la forma

$$F(n) = 2^{2^n} + 1$$

son primos. En efecto, para $n = 1, 2, 3, 4$ se obtiene

$$\begin{aligned} F(1) &= 2^2 + 1 = 5, \\ F(2) &= 2^{2^2} + 1 = 2^4 + 1 = 17, \\ F(3) &= 2^{2^3} + 1 = 2^8 + 1 = 257, \\ F(4) &= 2^{2^4} + 1 = 2^{16} + 1 = 65\,537, \end{aligned}$$

todos los cuales son números primos. Pero, en 1732, Euler obtuvo la descomposición $2^{2^5} + 1 = 641 \cdot 6\,700\,417$; de donde resulta que $F(5)$ no es primo. Más tarde se comprobó que otros de los «números de Fermat» eran compuestos, siendo necesarios para dichas comprobaciones métodos difíciles de la teoría de números, a causa de la insuperable dificultad de los ensayos directos. Hasta la fecha, no ha sido probado que sea primo ninguno de los números $F(n)$ para $n > 4$.

Otra sencilla y notable expresión que da varios números primos es

$$f(n) = n^2 - n + 41.$$

Para $n = 1, 2, ..., 40$ $f(n)$ es primo; pero para $n = 41$, se tiene $f(n) = 41^2$, que evidentemente no es primo.

La expresión

$$n^2 - 79n + 1601$$

da primos para n menor que 80, pero falla para $n = 80$. En resumen, puede decirse que la tarea de encontrar expresiones de tipo sencillo que den números primos ha resultado estéril. Ni que decir tiene que la posibilidad de obtener una fórmula algébrica simple que diera *todos* los números primos aparece como irrealizable.

b) *Números primos en una progresión aritmética.*—Mientras que es muy sencillo demostrar que existen infinitos números primos en la sucesión de enteros 1, 2, 3, ..., la extensión de un resultado análogo para sucesiones tales como la 1, 4, 7, 10, 13, ... o la 3, 7, 11, 15, 19, ... o, más en general, para toda progresión aritmética, $a, a + d, a + 2d, ...$ $a + nd, ...$, donde a y d no tienen factores comunes, es bastante más complicada. Todas las observaciones parecen indicar el hecho de que *en toda progresión de dicho tipo existen infinitos números primos*, como ocurre para la progresión más simple: 1, 2, 3, ... El probar este teorema general requirió un gran esfuerzo. Lejeune-Dirichlet (1805-1859),

uno de los matemáticos más notables del siglo pasado, obtuvo ese brillante resultado mediante la aplicación de los métodos más avanzados del análisis matemático de su tiempo. Sus trabajos originales sobre este tema se cuentan aún hoy entre los resultados más destacados y, pasado ya un siglo, la demostración no ha sido lo suficientemente simplificada como para ser accesible a quienes no posean una buena preparación en la técnica del cálculo y de la teoría de funciones.

Aunque no vamos a intentar probar el teorema de Dirichlet en toda su generalidad, es fácil generalizar la demostración de Euclides sobre la infinitud de la sucesión de números primos, para extenderla a algunas progresiones aritméticas *especiales*, tales como $4n + 3$ y $6n + 5$. Para tratar la primera, observemos que todo número primo mayor que 2 es impar (ya que en otro caso sería divisible por 2) y, por tanto, será de la forma $4n + 1$ ó $4n + 3$, para ciertos enteros n. Por otra parte, el producto de dos números de la forma $4n + 1$ es también de esta forma, ya que

$$(4a + 1)(4b + 1) = 16ab + 4a + 4b + 1 = 4(4ab + a + b) + 1.$$

Supongamos ahora que no existiera más que un número finito de números primos $p_1, p_2, p_3, ..., p_n$, de la forma $4n + 3$, y consideremos el número

$$N = 4(p_1p_2 \ldots p_n) - 1 = 4(p_1 \ldots p_n - 1) + 3.$$

O bien N es primo o puede ser descompuesto en producto de números primos, ninguno de los cuales puede ser $p_1, p_2, ..., p_n$, puesto que dividiendo N por cualquiera de éstos, da de resto -1. Además, no todos los factores de N pueden ser de la forma $4n + 1$, ya que N no es de esta forma y, como acabamos de ver, el producto de números de la forma $4n + 1$ es también de esta forma. Por tanto, al menos uno de dichos factores debe ser de la forma $4n + 3$, lo cual es imposible, puesto que hemos visto que ninguno puede ser de los p y éstos formaban, según nuestra hipótesis, *todos* los números primos de la forma $4n + 3$. En consecuencia, la hipótesis de que no hay más que un número finito de números primos de la forma $4n + 3$ nos ha llevado a una contradicción, lo que demuestra que dicho número debe ser infinito.

Ejercicio: Pruébese el correspondiente teorema para la progresión $6n + 5$.

c) *El teorema de los números primos.*—En el proceso de la investigación en busca de una ley que gobernase la distribución de los números primos, se dió un paso decisivo cuando los matemáticos, dejando de lado los fútiles intentos para hallar una fórmula simple que diera

todos los números primos o el número exacto de éstos comprendidos entre los n primeros enteros, buscaron en su lugar información relativa a la distribución *media* de los números primos dentro del conjunto de los enteros.

Para todo entero n, designemos con A_n el número de primos comprendidos entre los enteros 1, 2, 3, ..., n. Si subrayamos los números primos de la sucesión formada por los primeros enteros 1, $\underline{2}$, $\underline{3}$, 4, $\underline{5}$, 6, $\underline{7}$, 8, 9, 10, $\underline{11}$, 12, $\underline{13}$, 14, 15, 16, $\underline{17}$, 18, $\underline{19}$, ... podremos calcular los valores iniciales de A_n:

$$A_1 = 0,\ A_2 = 1,\ A_3 = A_4 = 2,\ A_5 = A_6 = 3,\ A_7 = A_8 = A_9 = A_{10} = 4,$$
$$A_{11} = A_{12} = 5,\ A_{13} = A_{14} = A_{15} = A_{16} = 6,\ A_{17} = A_{18} = 7,\ A_{19} = 8, \text{ etc.}$$

Si tomamos ahora una sucesión de valores para n que crezca ilimitadamente; p. ej.,

$$n = 10,\ 10^2,\ 10^3,\ 10^4,\ \ldots,$$

la correspondiente sucesión de valores de A_n,

$$A_{10},\ A_{10^2},\ A_{10^3},\ A_{10^4},\ \ldots,$$

crecerá también sin límite (aunque menos rápidamente). Puesto que sabemos que existen infinitos números primos, los valores de A_n excederán más pronto o más tarde a cualquier número finito. La *densidad* de los números primos entre los n primeros enteros vendrá dada por el cociente A_n/n, y a partir de la tabla de números primos, los valores de A_n/n pueden ser calculados empíricamente para un gran número de valores de n.

n	A_n/n
10^3	0,168
10^6	0,078 498
10^9	0,050 847 478
...	....

El último número decimal de esta tabla puede considerarse como la probabilidad para que un entero elegido al azar entre los 10^9 primeros enteros sea primo, ya que el número de casos posibles es 10^9, de los cuales A_{10^9} son primos.

La distribución de números primos particulares entre los enteros es muy irregular; pero esta irregularidad *en pequeño* desaparece si fijamos nuestra atención en la distribución media de los números primos dada por la razón A_n/n. La ley sencilla que rige el comportamiento de este cociente es uno de los más notables descubrimientos

de toda la matemática. Para establecer el *teorema de los números primos*, debemos previamente definir el «logaritmo natural» de un entero n. Con este objeto, tomemos en el plano dos ejes perpendiculares y consideremos el conjunto de los puntos del plano para los cuales el producto de sus distancias x, y a estos dos ejes es igual a la unidad. En función de las coordenadas x, y, este lugar, una hipérbola equilátera, queda definido por la ecuación $xy = 1$. El log n vendrá definido entonces como el área limitada, en la figura 5, por la hipérbola, el eje x y las dos verticales $x = 1$ y $x = n$. (Una exposición más detallada del logaritmo se dará en el capítulo VIII.) Del estudio empírico de la tabla de números primos, Gauss dedujo que el cociente A_n/n es aproximadamente igual a $1/\log n$, y que la aproximación mejora al crecer n. La bondad de la aproximación viene medida por el cociente $\dfrac{A_n/n}{1/\log n}$, cuyos valores para $n = 1000$, 1 000 000, 1 000 000 000 se dan en la tabla siguiente.

n	A_n/n	$1/\log n$	$\dfrac{A_n/n}{1/\log n}$
10^3	0,168	0,145	1,159
10^6	0,078 498	0,072 382	1,084
10^9	0,050 847 478	0,048 254 942	1,053
. . .	. . . .	. . . .	. . . .

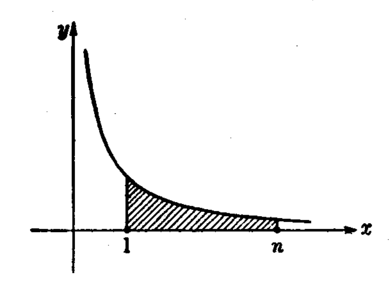


FIG. 5.—El área rayada por debajo de la hipérbola define log n.

Sobre la base de esta evidencia empírica, Gauss hizo la conjetura de que la razón A_n/n es «asintóticamente igual» a $1/\log n$. Lo que quiere decir que si tomamos una sucesión creciente de valores de n; p. ej., si tomamos para n

$$10,\ 10^2,\ 10^3,\ 10^4,\ \ldots$$

como antes hicimos, la razón de A_n/n a $1/\log n$,

$$\frac{A_n/n}{1/\log n},$$

calculada para estos valores sucesivos de n, se aproximará cada vez más a 1, y que la diferencia entre dicho cociente y 1 será tan pequeña como queramos para valores suficientemente grandes de n. Esta afirmación se expresa simbólicamente, mediante el signo $\sim$, en la forma:

$$\frac{A_n}{n} \sim \frac{1}{\log n}, \text{ lo que significa que } \frac{A_n/n}{1/\log n} \text{ tiende a 1 al crecer } n.$$

Que $\sim$ no puede ser reemplazado por el signo ordinario de igualdad $=$, resulta evidente si se observa que mientras A_n es siempre entero, $n/\log n$ no lo es.

El hecho de que el comportamiento medio de la distribución de los números primos pueda ser descrito mediante la función logarítmica es un descubrimiento notable, ya que resulta sorprendente que dos conceptos matemáticos que parecen tan inconexos estén en realidad tan íntimamente ligados.

Aunque el enunciado de la conjetura de Gauss es fácil de comprender, una demostración rigurosa de dicha conjetura era inaccesible a los métodos de la ciencia matemática de su época. Para probar el teorema, que se refiere únicamente a conceptos elementales, es necesario emplear los métodos más potentes de la matemática moderna. Fueron precisos casi cien años antes que el análisis se desarrollara lo suficiente para que Hadamard (1896), en París, y de la Vallée Poussin (1896), en Lovaina, pudieran dar una demostración completa del teorema de los números primos. Se han hecho simplificaciones y modificaciones importantes por v. Mangoldt y Landau. Mucho antes del resultado de Hadamard, Riemann (1826-1866) había contribuído al problema con una decisiva labor de exploración en un célebre trabajo, en el cual aparecen definidas las direcciones estratégicas del posible ataque. Recientemente, el matemático norteamericano Norbert Wiener ha sido capaz de modificar la demostración de modo que evita el uso de números complejos en un paso importante del razonamiento. En todo caso, la demostración del teorema de los números primos continúa siendo una cuestión difícil, aun para estudiantes avanzados. Volveremos a ocuparnos del tema en el capítulo VIII, suplemento IV.[1]

[1] En 1949 un joven matemático noruego, Selberg, y el matemático húngaro Erdös, en dos trabajos en cierto modo complementarios, dieron una demostración *elemental* (¡aunque no sencilla!) del teorema de los números primos. La demostración, que no depende de ideas analíticas ajenas al problema, constituye un descubrimiento de primordial importancia para la estructura lógica de la teoría de la distribución de los números primos. *(N. del T.)*

d) *Dos problemas no resueltos referentes a los números primos.*— Mientras que el problema de la distribución media de los números primos ha sido resuelto satisfactoriamente, quedan otras varias conjeturas que, sostenidas por la evidencia empírica, no han podido hasta ahora ser probadas como ciertas.

Una de ellas es la famosa *conjetura de Goldbach.* Goldbach (1690-1764) aparece en la historia de las matemáticas únicamente por esa conjetura que propuso como problema en una carta a Euler en 1742. Comprobó que en todos los casos observados, todo número par (excepto el 2, que es primo) puede ser representado como suma de dos números primos; p. ej.: $4 = 2 + 2, 6 = 3 + 3, 8 = 5 + 3, 10 = 5 + 5, 12 = 5 + 7, 14 = 7 + 7, 16 = 13 + 3, 18 = 11 + 7, 20 = 13 + 7, ..., 48 = 29 + 19, ..., 100 = 97 + 3$, etcétera.

Goldbach preguntaba a Euler si era capaz de demostrar que esa propiedad es cierta para *todo* número par, o si podría encontrar un contraejemplo. Euler no pudo dar una respuesta, ni nadie ha podido darla hasta ahora. La evidencia empírica a favor de la proposición de que todo número par puede ser representado de ese modo es bastante convincente, como puede verificar cualquiera ensayando en algunos ejemplos. El origen de la dificultad reside en el hecho de que los números primos se definen mediante la *multiplicación*, mientras que el problema se refiere a la *adición*. En términos generales, resulta difícil establecer conexiones entre las propiedades aditivas y las multiplicativas de los enteros.

Hasta hace poco tiempo, una demostración de la conjetura de Goldbach parecía completamente inaccesible. Hoy, en cambio, no parece estar muy lejos la solución de este problema. Un éxito importante, completamente inesperado y que sorprendió a los expertos, fué alcanzado en 1931 por un joven matemático ruso desconocido hasta entonces, Schnirelmann (1905-1938), quien probó que *todo entero positivo puede ser representado como suma de a lo sumo 300 000 números primos.* Aunque este resultado pueda parecer ridículo en comparación con el problema inicial de probar la conjetura de Goldbach, fué, sin embargo, el primer paso en la dirección justa. La prueba es directa y constructiva, aunque no da un método práctico para encontrar la descomposición en suma de números primos para un entero arbitrario. Posteriormente, el matemático ruso Vinogradoff, usando métodos debidos a Hardy, Littlewood y a su gran colaborador indio Ramanujan, ha conseguido reducir el número de 300 000 a 4, lo que está mucho más cerca de la solución del problema de Goldbach. Sin embargo, hay una diferencia notable entre los resultados de Schnirelmann y Vinogradoff,

más significativa aún que la diferencia entre 300 000 y 4. El teorema de Vinogradoff ha sido demostrado únicamente para enteros «suficientemente grandes»; con más precisión: Vinogradoff ha probado que *existe* un entero N tal que todo entero $n > N$ puede ser representado como suma de, a lo más, cuatro números primos. La demostración de Vinogradoff no permite determinar N; en contraste con el teorema de Schnirelmann, el de Vinogradoff es esencialmente indirecto y no constructivo. Lo que realmente prueba Vinogradoff es que la hipótesis de que existen infinitos enteros que no pueden descomponerse en suma de 4 números primos es absurda. Aquí se tiene un buen ejemplo de la profunda diferencia existente entre los dos tipos de demostración, directa e indirecta. (Véase la discusión general en la página 95.)

El problema siguiente, más notable aún que el de Goldbach, está todavía muy lejos de ser resuelto. Se ha observado que los números primos se suceden frecuentemente en pares de la forma p y $p + 2$. Así, p. ej., 3 y 5, 11 y 13, 29 y 31, etc. Se cree que la proposición que afirma la existencia de infinitos pares de ese tipo es cierta; sin embargo, hasta el presente no se ha podido dar el menor paso aprovechable en el camino de demostrar tal proposición.

II. CONGRUENCIAS

1. Conceptos generales.—Siempre que se presenta la cuestión de la divisibilidad de enteros por un entero fijo d, el concepto y la notación de *congruencia* (debidos a Gauss) sirven para aclarar y simplificar el razonamiento.

Para introducir este concepto, consideremos los restos que dan los sucesivos enteros al dividirlos por 5; se tiene

$0 = 0 \cdot 5 + 0$	$7 = 1 \cdot 5 + 2$	$-1 = -1 \cdot 5 + 4$
$1 = 0 \cdot 5 + 1$	$8 = 1 \cdot 5 + 3$	$-2 = -1 \cdot 5 + 3$
$2 = 0 \cdot 5 + 2$	$9 = 1 \cdot 5 + 4$	$-3 = -1 \cdot 5 + 2$
$3 = 0 \cdot 5 + 3$	$10 = 2 \cdot 5 + 0$	$-4 = -1 \cdot 5 + 1$
$4 = 0 \cdot 5 + 4$	$11 = 2 \cdot 5 + 1$	$-5 = -1 \cdot 5 + 0$
$5 = 1 \cdot 5 + 0$	$12 = 2 \cdot 5 + 2$	$-6 = -2 \cdot 5 + 4$
$6 = 1 \cdot 5 + 1$	etc.	etc.

Observamos que el resto de cualquier entero es uno de los números 0, 1, 2, 3, 4. Diremos que dos enteros a y b son «congruentes módulo 5» si ambos dan el *mismo resto* al ser divididos por 5. Así, 2, 7, 12 17, 22, ..., -3, -8, -13, -18, ..., son todos congruentes módulo 5, puesto que todos dan de resto 2. En general, diremos que dos enteros a y b son *congruentes módulo d*, siendo d un entero dado, si a y b dan

el mismo resto al dividirlos por d, lo que equivale a decir que hay un entero n tal que $a - b = nd$; p. ej., 27 y 15 son congruentes módulo 4, puesto que

$$27 = 6 \cdot 4 + 3, \qquad 15 = 3 \cdot 4 + 3.$$

El concepto de congruencia es tan útil, que resulta conveniente disponer de una notación sencilla para él. Escribiremos

$$a \equiv b \qquad (\text{mód } d)$$

para expresar que a y b son congruentes módulo d. En los casos en que no haya duda acerca de cuál sea el módulo, el «mód d» de la fórmula puede ser omitido. [Si a no es congruente con b módulo d, escribiremos $a \not\equiv b$ (mód d).]

Las congruencias aparecen frecuentemente en la vida diaria; p. ej., las manecillas de un reloj indican la hora módulo 12, y el cuentakilómetros de un coche da el total de kilómetros recorridos módulo 100 000.

Antes de proceder a una discusión detallada de las congruencias, el lector debe observar las siguientes proposiciones, que son todas ellas equivalentes:

1. a es congruente con b módulo d.
2. $a = b + nd$ para un cierto entero n.
3. d divide a $a - b$.

La utilidad de la notación de Gauss para las congruencias se basa en el hecho de que la congruencia respecto de un módulo fijo tiene varias de las propiedades formales de la igualdad ordinaria. Las propiedades formales más importantes de la relación $a = b$ son las siguientes:

1) Se tiene siempre $a = a$.
2) De $a = b$ se sigue $b = a$.
3) De $a = b$ y $b = c$ se sigue $a = c$.

Además, de $a = a'$ y $b = b'$ se sigue

4) $a + b = a' + b'$
5) $a - b = a' - b'$.
6) $ab = a'b'$.

Estas propiedades continúan siendo válidas si se reemplaza la relación $a = b$ por la de congruencia $a \equiv b$ (mód d). Así:

1') Se tiene siempre $a \equiv a$ (mód d).
2') De $a \equiv b$ (mód d) se sigue $b \equiv a$ (mód d).
3') De $a \equiv b$ (mód d) y $b \equiv c$ (mód d) se sigue $a \equiv c$ (mód d).

Dejamos al cuidado del lector el comprobar estos hechos.

Además, de $a \equiv a'$ (mód d) y $b \equiv b'$ (mód d) se sigue

4') $a + b \equiv a' + b'$ (mód d).
5') $a - b \equiv a' - b'$ (mód d).
6') $ab \equiv a'b'$ (mód d).

Resulta, en consecuencia, que varias *congruencias respecto del mismo módulo pueden sumarse, restarse* y *multiplicarse.* Para probar estas tres proposiciones es suficiente observar que si

$$a = a' + rd, \qquad b = b' + sd,$$

se tiene

$$a + b = a' + b' + (r + s)d,$$
$$a - b = a' - b' + (r - s)d,$$
$$ab = a'b' + (a's + b'r + rsd)d,$$

de donde resultan las conclusiones deseadas.

El concepto de congruencia tiene una interpretación geométrica muy intuitiva. De ordinario, cuando se quieren representar los enteros geométricamente, se elige un segmento de longitud unidad y se



FIG. 6.—Representación geométrica de los números enteros.

llevan éste y sus múltiplos en los dos sentidos. De este modo se hace corresponder a cada entero un punto sobre la recta, como en la figura 6. Si se trata de los enteros módulo d, dos números congruentes se consideran como uno mismo en lo que a la división por d se refiere, puesto que los dos dan el mismo resto. Para ver esto geométricamente, utilicemos una circunferencia dividida en d partes iguales. Cualquier entero dividido por d da de resto uno de los d números 0, 1, ..., $d - 1$, que están situados a intervalos iguales sobre la circunferencia. Cualquier entero es congruente módulo d con uno de estos d números, y, por tanto, puede ser representado geométricamente por uno de esos puntos; dos números son congruentes si están representados por el mismo punto. La figura 7 está dibujada para el caso $d = 6$. La esfera de un reloj es otro ejemplo tomado de la vida corriente.

Como aplicación de la propiedad multiplicativa 6') de las congruencias, determinaremos los restos de las sucesivas potencias de 10 respecto de un número dado; p. ej.,

$$10 \equiv -1 \text{ (mód. 11)},$$

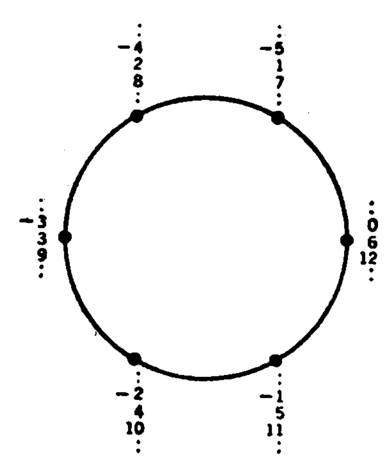


FIG. 7.—Representación geométrica de los números enteros respecto al módulo 6.

ya que $10 = -1 + 11$. Por multiplicaciones sucesivas de esta congruencia por sí misma se obtiene

$$10^2 \equiv (-1)(-1) = 1 \quad (\text{mód } 11),$$
$$10^3 \equiv -1 \quad \text{»} \quad ,$$
$$10^4 \equiv 1 \quad \text{»} \quad , \text{ etc.}$$

Como consecuencia, se puede probar que cualquier entero

$$z = a_0 + a_1 \cdot 10 + a_2 \cdot 10^2 + \dots + a_n \cdot 10^n,$$

expresado en el sistema decimal, da al dividirlo por 11 el mismo resto que la suma de sus cifras, tomadas alternativamente con los signos más y menos,

$$t = a_0 - a_1 + a_2 - a_3 + \dots .$$

En efecto, se puede escribir

$$z - t = a_1 \cdot 11 + a_2(10^2 - 1) + a_3(10^3 + 1) + a_4(10^4 - 1) + \cdots$$

Y como todos los números 11, $10^2 - 1$, $10^3 + 1$, ... son congruentes con 0 módulo 11, también lo será $z - t$; por consiguiente, z da el mismo resto que t al dividirlo por 11. Resulta en particular que un número es divisible por 11 (es decir, da de resto 0) cuando, y únicamente entonces, la suma alternada de sus cifras es divisible por 11. Por ejemplo, puesto que $3 - 1 + 6 - 2 + 8 - 1 + 9 = 22$, el número $z =$ $= 3\,162\,819$ es divisible por 11. Encontrar una regla de divisibilidad

por 3 ó por 9 es más sencillo aún, puesto que $10 \equiv 1$ (mód 3 ó 9), y, por consiguiente, $10^n \equiv 1$ (mód 3 ó 9) para todo n. Resulta así que para que un número z sea divisible por 3 ó por 9 es necesario y suficiente que la suma de sus cifras

$$s = a_0 + a_1 + a_2 + \ldots + a_n$$

sea también divisible por 3 ó por 9, respectivamente.

Para las congruencias módulo 7 se tiene

$$10 \equiv 3, \quad 10^2 \equiv 2, \quad 10^3 \equiv -1, \quad 10^4 \equiv -3, \quad 10^5 \equiv -2, \quad 10^6 \equiv 1.$$

Los restos sucesivos se repiten, de donde resulta que para que z sea divisible por 7 es necesario y suficiente que la expresión

$$r = a_0 + 3a_1 + 2a_2 - a_3 - 3a_4 - 2a_5 + a_6 + 3a_7 + \ldots$$

sea divisible por 7.

Ejercicio: Hállese una regla análoga para la divisibilidad por 13.

Sumando o multiplicando congruencias respecto de un módulo fijo, p. ej., $d = 5$, podemos evitar los números muy grandes al reemplazar cualquier número a por el del conjunto.

$$0, \ 1, \ 2, \ 3, \ 4$$

congruente con él. Así, para calcular sumas y productos módulo 5 necesitamos únicamente las siguientes tablas de adición y multiplicación:

$a + b$

	$b \equiv 0$	1	2	3	4
$a \equiv 0$	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$a \cdot b$

	$b \equiv 0$	1	2	3	4
$a \equiv 0$	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

De la segunda de esas tablas resulta que un producto ab es congruente con 0 (mód 5) únicamente si a o b son $\equiv 0$ (mód 5). Esto sugiere la ley general

7) $ab \equiv 0$ (mód d) únicamente si $a \equiv 0$ ó $b \equiv 0$ (mód d),

la cual resultaría como una extensión de la ley ordinaria para enteros que dice que se tiene $ab = 0$ únicamente si es $a = 0$ ó $b = 0$. Ahora bien: *la ley* 7) *es válida solamente cuando el módulo* d *es un número primo*. En efecto, la congruencia

$$ab \equiv 0 \qquad (\text{mód } d)$$

significa que d divide a ab, y hemos visto antes que un número primo divide a un producto ab únicamente si divide a a o a b; esto es, únicamente si

$$a \equiv 0 \pmod{d} \quad \text{ó} \quad b \equiv 0 \pmod{d}.$$

Si d no es primo, esta regla no vale en general; puesto que si es $d = r \cdot s$, siendo r y s menores que d, se tiene

$$r \not\equiv 0 \pmod{d}, \qquad s \not\equiv 0 \pmod{d},$$

y, en cambio, es

$$rs = d \equiv 0 \pmod{d}.$$

Por ejemplo, $2 \not\equiv 0 \pmod{6}$ y $3 \not\equiv 0 \pmod{6}$; sin embargo, $2 \cdot 3 = 6 \equiv 0 \pmod{6}$.

Ejercicio: Pruébese que la siguiente *regla de simplificación* es válida para congruencias respecto de un módulo primo:

Si es $ab \equiv ac$ y $a \not\equiv 0$, se tiene $b \equiv c$.

Ejercicios:

1. ¿Con qué número entre 0 y 6 inclusive es congruente módulo 7 el producto $11 \cdot 18 \cdot 2322 \cdot 13 \cdot 19$?

2. ¿Con qué número comprendido entre 0 y 12 inclusive es congruente módulo 13 el producto $3 \cdot 7 \cdot 11 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 113$?

3. ¿Con qué número entre 0 y 4 inclusive es congruente módulo 5 la suma $1 + 2 + 2^2 + \ldots + 2^{10}$?

2. Teorema de Fermat.—En el siglo XVII, Fermat, el fundador de la moderna teoría de números, descubrió un teorema muy importante: *Si* p *es un número primo que no divide al entero* a, *se tiene*:

$$a^{p-1} \equiv 1 \pmod{p},$$

lo que significa que la potencia $(p - 1)$-ésima de a da resto 1 al dividirla por p.

Algunos de nuestros cálculos anteriores confirman este teorema; p. ej., encontrábamos que $10^6 \equiv 1 \pmod{7}$, $10^2 \equiv 1 \pmod{3}$, y $10^{10} \equiv 1 \pmod{11}$. Del mismo modo se vería que es $2^{12} \equiv 1 \pmod{13}$ y $5^{10} \equiv 1 \pmod{11}$. Para comprobar las últimas congruencias no es necesario calcular efectivamente las potencias indicadas, ya que pueden utilizarse ventajosamente las propiedades multiplicativas de aquéllas:

$2^4 = 16 \equiv 3$	(mód 13)	$5^2 \equiv 3$	(mód 11)
$2^8 \equiv 9 \equiv -4$	»	$5^4 \equiv 9 \equiv -2$	»
$2^{12} \equiv -4 \cdot 3 = -12 \equiv 1$	»	$5^8 \equiv 4$	»
		$5^{10} \equiv 3 \cdot 4 = 12 \equiv 1$	»

Para demostrar el teorema de Fermat, consideremos los múltiplos de a:

$$m_1 = a, \quad m_2 = 2a, \quad m_3 = 3a, \ldots, m_{p-1} = (p-1)a$$

Ningún par de estos enteros puede estar formado por números congruentes módulo p; de lo contrario, p sería un factor de $m_r - m_s = (r - s)a$ para algún par r, s, siendo $1 \leqslant s < r \leqslant (p - 1)$. Pero la regla 7) indica que esto no puede ocurrir; puesto que $r - s$ es menor que p, p no puede ser factor de $(r - s)$, y habíamos supuesto que p no dividía a a. Asimismo, ninguno de aquellos múltiplos puede ser congruente con 0. Por consiguiente, los números $m_1, m_2, m_3, \ldots, m_{p-1}$ deben ser congruentes con los números $1, 2, 3, \ldots, p - 1$ tomados en orden conveniente. Resulta como consecuencia

$$m_1 m_2 \cdots m_{p-1} = 1 \cdot 2 \cdot 3 \cdots (p-1)a^{p-1} \equiv 1 \cdot 2 \cdot 3 \cdots (p-1) \quad (\text{mód } p),$$

o, si escribimos por brevedad K en vez de $1 \cdot 2 \cdot 3 \ldots (p - 1)$,

$$K(a^{p-1} - 1) \equiv 0 \qquad (\text{mód } p).$$

Pero K no es divisible por p, ya que no lo es ninguno de sus factores; por tanto, en virtud de la regla 7), $(a^{p-1} - 1)$ debe ser divisible por p; es decir,

$$a^{p-1} - 1 \equiv 0 \qquad (\text{mód } p),$$

que es el teorema de Fermat.

Para comprobar el teorema una vez más, tomemos $p = 23$ y $a = 5$. Se tiene entonces (mód 23) $5^2 \equiv 2$, $5^4 \equiv 4$, $5^8 \equiv 16 \equiv -7$, $5^{16} \equiv 49 \equiv 3$, $5^{20} \equiv 12$, $5^{22} \equiv 24 \equiv 1$. Con $a = 4$, en vez de 5, se obtendría, también (mód 23), $4^2 \equiv -7$, $4^3 \equiv -28 \equiv -5$, $4^4 \equiv -20 \equiv 3$, $4^8 \equiv 9$, $4^{11} \equiv -45 \equiv 1$, $4^{22} \equiv 1$.

En el ejemplo anterior con $a = 4$, $p = 23$, y en otros, se observa que no solamente la potencia $(p - 1)$-ésima de a, sino también otras potencias menores pueden ser congruentes con 1. Pero ocurre siempre que la menor de dichas potencias, en dicho caso 11, es un divisor de $p - 1$. (Véase el ejercicio 3 siguiente.)

Ejercicios:

1. Compruébese mediante cálculos análogos a los anteriores que $2^8 \equiv 1$ (mód 17); $3^8 \equiv -1$ (mód 17); $3^{14} \equiv -1$ (mód 29); $2^{14} \equiv -1$ (mód 29); $4^{14} \equiv 1$ (mód 29); $5^{14} \equiv 1$ (mód 29).

2. Compruébese el teorema de Fermat para $p = 5, 7, 11, 17$ y 23, con diferentes valores de a.

3. Demuéstrese el teorema general siguiente: El menor entero positivo e para

el que se tenga $a^e \equiv 1$ (mód p), siendo p primo, debe ser un divisor de $p - 1$ [*Indicación*: Dividiendo $p - 1$ por e, se obtendría

$$p - 1 = ke + r,$$

siendo $0 < r < e$; utilícese el hecho de que $a^{p-1} \equiv a^e \equiv 1$ (mód p).]

3. Restos cuadráticos.—Por los ejemplos considerados en relación con el teorema de Fermat, se ha visto que no solamente es siempre $a^{p-1} \equiv 1$ (mód p), sino que (si p es un número primo distinto de 2, y, por consiguiente, impar y de la forma $p = 2p' + 1$) para algunos valores de a, $a^{p'} = a^{(p-1)/2} \equiv 1$ (mód p). Este hecho sugiere otras varias cuestiones interesantes. Podemos escribir el teorema en la siguiente forma:

$$a^{p-1} - 1 = a^{2p'} - 1 = (a^{p'} - 1)(a^{p'} + 1) \equiv 0 \qquad (\text{mód } p).$$

Puesto que un producto es divisible por p cuando uno al menos de los factores lo es, resulta inmediatamente que, bien $a^{p'} - 1$ o $a^{p'} + 1$ debe ser divisible por p, de modo que para cualquier primo $p > 2$ y cualquier número a no divisible por p, se tiene

$$a^{(p-1)/2} \equiv 1 \quad \text{o} \quad a^{(p-1)/2} \equiv -1 \qquad (\text{mód } p).$$

Desde el comienzo de la moderna teoría de números los matemáticos se han interesado por el problema de determinar qué números a pertenecen a la primera clase y cuáles a la segunda. Supongamos que a es congruente módulo p con el cuadrado de otro entero x,

$$a \equiv x^2 \qquad (\text{mód } p).$$

Entonces se tendrá: $a^{(p-1)/2} \equiv x^{p-1}$ y, de acuerdo con el teorema de Fermat, congruente con 1 módulo p. Un número a, no múltiplo de p, que sea congruente módulo p con el cuadrado de otro entero se llama *resto cuadrático de p*, mientras que un número b, no múltiplo de p, que no sea congruente con ningún cuadrado, se llama *no-resto cuadrático de p*. Acabamos de ver que todo resto cuadrático a de p satisface la congruencia $a^{(p-1)/2} \equiv 1$ (mód p). Sin grandes dificultades se puede probar que para cualquier no-resto b se tiene la congruencia $b^{(p-1)/2} \equiv -1$ (mód p). Por otra parte, vamos a demostrar que entre los números 1, 2, 3, ..., $p - 1$ hay precisamente $(p - 1)/2$ restos cuadráticos y $(p - 1)/2$ no-restos.

Aun siendo posible obtener muchos datos empíricos por cálculo directo, al comienzo no fué fácil descubrir leyes generales que rigieran la distribución de los restos y no-restos cuadráticos. La primera propiedad importante de los restos fué observada por Legendre (1752-

1833), y designada más tarde por Gauss con el nombre de *ley de reciprocidad cuadrática*. Esta ley se refiere al comportamiento de dos números primos distintos p y q, y dice que q es un resto cuadrático de p cuando, y sólo entonces, p es resto cuadrático de q, si el producto $[(p-1)/2]\,[(q-1)/2]$ es *par*. En el caso en que dicho producto sea *impar*, la situación es la inversa, de modo que p es resto de q si q es *no-resto* de p. Uno de los descubrimientos juveniles de Gauss fué la primera demostración rigurosa de este teorema, que durante mucho tiempo había desafiado al mundo matemático. La primera demostración de Gauss no era sencilla, y aún hoy la ley de reciprocidad no es fácil de establecer, a pesar de que se han dado de ella numerosas demostraciones. Su verdadera significación aparece claramente cuando se establece su conexión con los descubrimientos modernos en la teoría de los números algebraicos.

Como ejemplo para ilustrar la distribución de los restos cuadráticos, tomemos $p = 7$. Entonces, puesto que se tiene

$$0^2 \equiv 0, \quad 1^2 \equiv 1, \quad 2^2 \equiv 4, \quad 3^2 \equiv 2, \quad 4^2 \equiv 2, \quad 5^2 \equiv 4, \quad 6^2 \equiv 1,$$

todas módulo 7, y ya que los cuadrados restantes repiten esta sucesión, los restos cuadráticos de 7 son los números congruentes con 1, 2 ó 4, mientras que los no-restos son congruentes con 3, 5 ó 6. En el caso general, los restos cuadráticos de p son los números congruentes con $1^2, 2^2, \ldots, (p-1)^2$. Pero éstos son congruentes a pares, pues se tiene

$$x^2 \equiv (p-x)^2 \pmod{p} \qquad [\text{p. ej.}, 2^2 \equiv 5^2 \pmod{7}],$$

por ser $(p-x)^2 = p^2 - 2px + x^2 \equiv x^2 \pmod{p}$. Resulta así que la mitad de los números $1, 2, \ldots, p-1$ son restos cuadráticos de p y la otra mitad son no-restos cuadráticos.

Como ejemplo de la ley de reciprocidad tomemos $p = 5$, $q = 11$. Por ser $11 \equiv 1^2$ (mód 5), 11 es resto cuadrático (mód 5); y como el producto $[(5-1)/2]\,[(11-1)/2]$ es par, la ley de reciprocidad dice que 5 es resto cuadrático (mód 11). En confirmación de esto, observemos que se tiene $5 \equiv 4^2$ (mód 11). Por otra parte, si $p = 7$, $q = 11$, el producto $[(7-1)/2]\,[(11-1)/2]$ es impar, y en efecto 11 es un resto cuadrático (mód 7) [ya que $11 \equiv 2^2$ (mód 7)], mientras que 7 es no-resto cuadrático (mód 11).

Ejercicios:

1. $6^2 = 36 \equiv 13$ (mód 23). ¿Es 23 resto cuadrático (mód 13)?
2. Hemos visto que $x^2 \equiv (p-x)^2 \pmod{p}$. Demuéstrese que ésas son las *únicas* congruencias módulo p entre los números $1^2, 2^2, 3^2, \ldots, (p-1)^2$.

III. LOS NÚMEROS PITAGÓRICOS Y EL ÚLTIMO TEOREMA DE FERMAT

Una interesante cuestión de la teoría de números se halla relacionada con el teorema de Pitágoras. Los griegos sabían que un triángulo con lados 3, 4, 5 es un triángulo rectángulo. Se presentaba entonces la cuestión general: «¿Qué otros triángulos rectángulos tienen sus lados iguales a múltiplos enteros de la unidad de longitud?» El teorema de Pitágoras se expresa algébricamente por la ecuación

$$a^2 + b^2 = c^2, \qquad [1]$$

donde a y b son las longitudes de los catetos, y c, la de la hipotenusa. El problema de hallar *todos* los triángulos rectángulos con lados de longitud entera es equivalente al de hallar las soluciones enteras de la ecuación [1]. Una terna de dichos números se llama *terna de números pitagóricos.*

El problema de hallar todas las ternas de números pitagóricos puede resolverse fácilmente. Si a, b y c forman una terna pitagórica, de modo que $a^2 + b^2 = c^2$, pongamos para simplificar $a/c = x$, $b/c=y$, donde x e y son números racionales para los cuales se verifica $x^2 + y^2 = 1$. Entonces se tiene $y^2 = (1 - x)(1 + x)$, o lo que es lo mismo $y/(1 + x) = (1 - x)/y$. El valor común de los dos miembros de esta ecuación es un número t que puede expresarse como cociente de dos números enteros u/v. Podemos escribir entonces: $y = t(1 + x)$ y $(1 - x) = ty$; es decir,

$$tx - y = -t, \qquad x + ty = 1.$$

De este sistema de ecuaciones se deduce inmediatamente

$$x = \frac{1 - t^2}{1 + t^2}, \qquad y = \frac{2t}{1 + t^2}$$

y sustituyendo x, y, t por sus valores, se tiene

$$\frac{a}{c} = \frac{v^2 - u^2}{u^2 + v^2}, \qquad \frac{b}{c} = \frac{2uv}{u^2 + v^2}$$

Por consiguiente,

$$\begin{aligned} a &= (v^2 - u^2)r, \\ b &= (2uv)r, \\ c &= (u^2 + v^2)r, \end{aligned} \qquad [2]$$

con un factor arbitrario de proporcionalidad r. Esto prueba que si (a, b, c) es una terna pitagórica, a, b, c son proporcionales a $v^2 - u^2$,

$2uv$, $u^2 + v^2$, respectivamente. Recíprocamente, es fácil ver que toda terna (a, b, c) definida mediante [2] es pitagórica, ya que de [2] se sigue

$$a^2 = (u^4 - 2u^2v^2 + v^4)r^2,$$
$$b^2 = (4u^2v^2)r^2,$$
$$c^2 = (u^4 + 2u^2v^2 + v^4)r^2,$$

de forma que $a^2 + b^2 = c^2$.

El resultado anterior puede simplificarse aún: a partir de cualquier terna pitagórica (a, b, c) se pueden obtener otras ternas pitagóricas (sa, sb, sc) para todo valor entero de s; así, de (3, 4, 5) se obtienen las (6, 8, 10), (9, 12, 15), etc. Tales ternas no son esencialmente distintas, ya que corresponden a triángulos rectángulos semejantes. Definiremos como *primitiva* una terna pitagórica cuando los números a, b y c no tengan ningún factor común. Se puede probar que las *fórmulas*

$$a = v^2 - u^2,$$
$$b = 2uv,$$
$$c = u^2 + v^2,$$

dan para todo par de enteros positivos u, v, *con* u $>$ v, *y tales que* u *y* v *no tengan factores comunes ni sean los dos impares, todas las ternas primitivas de números pitagóricos.*

***Ejercicio:** Demuéstrese la última proposición.

Como ejemplos de ternas primitivas de números pitagóricos se tiene: $u = 2$, $v = 1$: (3, 4, 5); $u = 3$, $v = 2$: (5, 12, 13); $u = 4$, $v = 3$: (7, 24, 25); ...; $u = 10$, $v = 7$: (51, 140; 149); etc.

Los resultados alcanzados para los números pitagóricos inducen de modo natural a plantear la cuestión de obtener enteros a, b, c para los que se verifique $a^3 + b^3 = c^3$ o $a^4 + b^4 = c^4$, o, en general, para un exponente entero positivo $n > 2$ dado, a determinar las soluciones enteras y positivas de la ecuación

$$a^n + b^n = c^n. \qquad [3]$$

Fermat dió a esta cuestión una respuesta de modo un poco espectacular. Estudiando la obra de Diofanto, el que más había contribuído entre los antiguos a la teoría de números, Fermat tenía la costumbre de escribir comentarios en las márgenes del libro. De esta forma enunció sin demostración muchos teoremas. Todos han sido probados posteriormente, con excepción de uno de ellos de especial importancia. Comentando la teoría de los números pitagóricos, Fermat escribió que

la ecuación [3] *no admite soluciones enteras para cualquier* n > 2, pero que la elegante demostración que había encontrado era desgraciadamente demasiado larga para el margen de que disponía.

De esta proposición general no han podido ser demostradas ni su validez ni su falsedad a pesar de los esfuerzos de muchos de los más grandes matemáticos posteriores a Fermat. El teorema ha sido probado para varios valores de n; en particular para todo $n < 619$, pero no para todo n, sin haberse encontrado hasta el presente ningún contraejemplo. Aunque el teorema en sí no es de gran importancia matemática, las tentativas hechas para demostrarlo han dado lugar a importantes investigaciones en la teoría de números. Este problema ha despertado también gran interés en círculos no matemáticos, debido en parte a un premio de 100 000 marcos ofrecido a la primera persona que diera una solución aceptada por la Real Academia de Gotinga. Hasta que la inflación que siguió en Alemania a la primera guerra mundial quitó todo valor económico a dicho premio, gran número de «soluciones» incorrectas eran enviadas todos los años a la Academia. Incluso matemáticos serios han creído a veces haber encontrado e incluso han publicado soluciones «correctas»; sin embargo, hasta el presente en todas ellas ha sido descubierto algún error. Con la devaluación del marco, el interés general parece haber desaparecido, aunque de cuando en cuando la prensa anuncia que el problema ha sido resuelto por un genio hasta entonces desconocido.

IV. EL ALGORITMO DE EUCLIDES

1. Teoría general.—Es bien conocida la regla de división de un entero a por otro b, y el lector sabe que el proceso de división no termina hasta que se llega a un resto más pequeño que el divisor. Así, si $a = 648$ y $b = 7$ se obtiene un cociente $q = 92$ y un resto $r = 4$.

```
648 | 7
63  |----
--- | 92
 18 |
 14 |
 -- |
  4 |
```

$$648 = 7 \cdot 92 + 4.$$

Lo que se puede enunciar como un teorema general: *Si* a *es un entero cualquiera y* b *es un entero mayor que* 0, *se pueden encontrar siempre dos enteros* q *y* r *tales que*

$$a = b \cdot q + r, \qquad [1]$$
$$0 \leqslant r < b.$$

Para probar esta proposición sin utilizar la regla ordinaria de división de enteros, basta observar que cualquier entero a es o bien múltiplo de b,

$$a = bq,$$

o está comprendido entre dos múltiplos consecutivos de b,

$$bq < a < b(q + 1) = bq + b.$$

En el primer caso se tiene [1] con $r = 0$, y en el segundo resulta, de la primera desigualdad anterior,

$$a - bq = r > 0,$$

mientras que la segunda desigualdad da

$$a - bq = r < b,$$

de modo que resulta $0 < r < b$, como se dice en [1].

De este sencillo hecho se deduce una gran variedad de consecuencias importantes; la primera de ellas es un método para hallar el máximo común divisor de dos enteros.

Sean a y b dos enteros cualesquiera, ninguno de los cuales sea 0, y consideremos el conjunto de todos los enteros positivos que dividen simultáneamente a a y b. Este conjunto es evidentemente finito, puesto que si, p. ej., es $a \neq 0$, ningún entero mayor que a puede ser divisor de a. En consecuencia, no puede haber más que un número finito de divisores comunes a a y b, y entre ellos habrá uno mayor que todos, d. Este número entero d se llama *máximo común divisor* de a y b, y se escribe $d = (a, b)$. Así, para $a = 8$ y $b = 12$ se obtiene por ensayos directos $(8, 12) = 4$, mientras que para $a = 5$ y $b = 9$ resulta $(5, 9) = 1$. Si a y b son números grandes, p. ej., $a = 1804$ y $b = 328$, el método de ensayos sería muy laborioso e incierto. Un método breve y seguro lo da el llamado *algoritmo de Euclides*. (Un algoritmo es un método sistemático de cálculo.) Está basado en el hecho de que de toda relación de la forma

$$a = b \cdot q + r \qquad [2]$$

se sigue

$$(a, b) = (b, r), \qquad [3]$$

puesto que todo número u que divida a a y b

$$a = su, \qquad b = tu,$$

divide también a r, ya que es $r = a - bq = su - qtu = (s - qt)u$; y recíprocamente, todo número v que divide a b y r,

$$b = s'v, \qquad r = t'v.$$

divide también a a, ya que es $a = bq + r = s'vq + t'v = (s'q + t')v$. Por consiguiente, *todo* divisor común de a y b es también divisor común de b y r, y recíprocamente. Por tanto, siendo el conjunto de *todos* los divisores comunes a a y b idéntico con el conjunto de *todos* los divisores comunes a b y r, el *mayor* de los divisores comunes a a y b debe ser igual al mayor de los divisores comunes a b y r, lo que demuestra [3]. Veremos inmediatamente la utilidad de esta relación.

Volvamos a la cuestión de hallar el máximo común divisor de 1804 y 328. Por división entera obtenemos

$$\begin{array}{r|l} 1804 & 328 \\ \cline{2-2} 1640 & 5 \\ \hline 164 & \end{array}$$

de donde resulta

$$1804 = 5 \cdot 328 + 164.$$

De aquí se sigue, en virtud de [3],

$$(1804,\ 328) = (328,\ 164).$$

Se observa que el problema de hallar (1804, 328) ha sido reducido a otro análogo que se refiere a números más pequeños. Continuemos el proceso; de

$$\begin{array}{r|l} 328 & 164 \\ \cline{2-2} 328 & 2 \\ \hline 0 & \end{array}$$

resulta $328 = 2 \cdot 164 + 0$, de modo que $(328, 164) = (164, 0) = 164$. En consecuencia, $(1804, 328) = (328, 164) = (164, 0) = 164$, que es el resultado buscado.

Este proceso para hallar el máximo común divisor de dos números está expuesto en forma geométrica en los *Elementos* de Euclides. Para enteros arbitrarios a y b, no simultáneamente nulos, dicho proceso puede ser descrito aritméticamente en los términos siguientes.

Podemos suponer $b \neq 0$, puesto que $(a, 0) = a$. Entonces, por divisiones sucesivas escribiremos:

$$\begin{array}{ll} a = bq_1 + r_1 & (0 < r_1 < b) \\ b = r_1q_2 + r_2 & (0 < r_2 < r_1) \\ r_1 = r_2q_3 + r_3 & (0 < r_3 < r_2) \\ r_2 = r_3q_4 + r_4 & (0 < r_4 < r_3) \\ \dots\dots\dots\dots & \dots\dots\dots\dots \end{array} \qquad [4]$$

mientras los restos $r_1, r_2, r_3, \ldots$ son distintos de 0. De las desigualdades que aparecen a la derecha en [4] resulta que los restos sucesivos forman una sucesión decreciente de números positivos:

$$b > r_1 > r_2 > r_3 > r_4 > \cdots > 0. \qquad [5]$$

Por tanto, a lo sumo al cabo de b divisiones (bastante menos, ya que la diferencia entre dos restos sucesivos es en general mayor que 1), se debe llegar al resto 0:

$$r_{n-2} = r_{n-1}q_n + r_n$$
$$r_{n-1} = r_n q_{n+1} + 0.$$

Cuando esto ocurre, sabemos que es

$$(a, b) = r_n ;$$

en otros términos, (a, b) es *igual al último resto mayor que* 0 *en la sucesión* [5]. Esto se deduce de la aplicación sucesiva de la igualdad [3] a las ecuaciones [4], puesto que de las igualdades de [4] resulta

$$(a, b) = (b, r_1), \qquad (b, r_1) = (r_1, r_2), \qquad (r_1, r_2) = (r_2, r_3)$$
$$(r_2, r_3) = (r_3, r_4), \ldots, (r_{n-1}, r_n) = (r_n, 0) = r_n.$$

Ejercicio: Aplíquese el algoritmo de Euclides para hallar el máximo común divisor de: *a)* 187, 77; *b)* 105, 385; *c)* 245, 193.

A partir de las ecuaciones [4] puede obtenerse una importante propiedad de (a, b). *Si es* $\mathrm{d} = (\mathrm{a}, \mathrm{b})$, *existen dos enteros, positivos o negativos*, k *y* l *tales que*

$$d = ka + lb. \qquad [6]$$

Para probarlo, consideremos la sucesión [5] de restos. De la primera ecuación en [4] resulta

$$r_1 = a - q_1 b,$$

de modo que r_1 puede escribirse en la forma $k_1 a + l_1 b$ (en este caso $k_1 = 1$, $l_1 = -q_1$). De la ecuación siguiente en [4] se obtiene

$$r_2 = b - q_2 r_1 = b - q_2(k_1 a + l_1 b) = (-q_2 k_1)a + (1 - q_2 l_1)b = k_2 a + l_2 b.$$

Evidentemente, este proceso puede repetirse para los restos sucesivos $r_3, r_4, \ldots$ hasta llegar a una relación

$$r_n = ka + lb,$$

como queríamos probar.

Como ejemplo, consideremos el algoritmo de Euclides para hallar (61, 24); el máximo común divisor es 1, y la representación deseada para 1 puede calcularse a partir de las ecuaciones

$$61 = 2 \cdot 24 + 13; \quad 24 = 1 \cdot 13 + 11; \quad 13 = 1 \cdot 11 + 2;$$
$$11 = 5 \cdot 2 + 1; \quad 2 = 2 \cdot 1 + 0.$$

De la primera de estas ecuaciones obtenemos

$$13 = 61 - 2 \cdot 24;$$

de la segunda,

$$11 = 24 - 13 = 24 - (61 - 2 \cdot 24) = -61 + 3 \cdot 24;$$

de la tercera,

$$2 = 13 - 11 = (61 - 2 \cdot 24) - (-61 + 3 \cdot 24) = 2 \cdot 61 - 5 \cdot 24,$$

y de la cuarta,

$$1 = 11 - 5 \cdot 2 = (-61 + 3 \cdot 24) - 5(2 \cdot 61 - 5 \cdot 24) = -11 \cdot 61 + 28 \cdot 24.$$

2. Aplicación al teorema fundamental de la aritmética.—El hecho de que $d = (a, b)$ pueda escribirse siempre en la forma $d = ka + lb$ puede utilizarse para dar una demostración del teorema fundamental de la aritmética, independiente de la dada en la página 30. Primero probaremos, como lema, el corolario de la página 31, y luego, a partir de este lema, deduciremos el teorema fundamental, invirtiendo así el orden de la demostración anterior.

LEMA: *Si un número primo* p *divide a un producto* ab, *divide a uno de los factores* a, b.

Si p no divide a a, por ser primo p, se debe tener $(a, p) = 1$, ya que los únicos divisores de p son p y 1. Por consiguiente, se podrán encontrar dos enteros k y l tales que

$$1 = ka + lp.$$

Multiplicando los dos miembros de esta igualdad por b, se obtiene

$$b = kab + lpb.$$

Ahora bien: si p divide a ab, se puede escribir

$$ab = pr$$

de modo que

$$b = kpr + lpb = p(kr + lb),$$

de donde resulta evidente que p divide a b. Hemos demostrado así que si p divide a ab y no divide a a, debe dividir a b; por consiguiente, en todo caso p dividirá a a o a b si divide a ab.

La extensión de este resultado a productos de más de dos factores es inmediata; p. ej., si p divide a abc, aplicando dos veces el lema podemos demostrar que p debe dividir al menos a uno de los enteros a, b o c. Ya que si p no divide ni a a, ni a b, ni a c, no puede dividir a ab y, en consecuencia, tampoco puede dividir a $(ab)c = abc$.

Ejercicio: La extensión de este razonamiento a productos de cualquier número n de enteros requiere el uso tácito o expreso del principio de inducción matemática. Complétense los detalles de tal razonamiento.

Del resultado anterior se sigue de modo inmediato el teorema fundamental de la aritmética. Supongamos que se tienen dos descomposiciones de un entero N en producto de números primos:

$$N = p_1p_2 \ldots p_r = q_1q_2 \ldots q_s .$$

Puesto que p_1 divide al segundo miembro de estas igualdades, debe dividir también al tercero, y, por tanto, por el ejercicio anterior, debe dividir a uno de los factores q_k. Pero q_k es primo; por consiguiente, p_1 debe ser igual a q_k. Suprimiendo este factor común en los dos últimos miembros, resultará que p_2 debe dividir a uno de los q_t restantes, y en consecuencia ser igual a él. Suprimiendo p_2 y q_t, procederíamos análogamente con p_3, ..., p_r. Al final de este proceso se habrán suprimido todas las p, dejando solamente la unidad en el segundo miembro. En el último miembro no podrá quedar ninguna q, puesto que todas las q son mayores que uno. Por consiguiente, las p y las q aparecen en parejas de números iguales, lo que prueba que, salvo quizá el orden de los factores, las dos descomposiciones son idénticas.

3. La función φ de Euler. De nuevo el teorema de Fermat.—Dos enteros a y b se llaman *primos relativos* cuando su máximo común divisor es 1:

$$(a, b) = 1.$$

Por ej., 24 y 35 son primos relativos, mientras que 12 y 18 no lo son. *Si* a *y* b *son primos relativos, existen dos enteros positivos o negativos* k *y* l *tales que*

$$ka + lb = 1.$$

Esto se sigue de la propiedad de (a, b) establecida en la página 53.

Ejercicio: Demuéstrese el teorema siguiente: *Si un entero* r *divide a un producto* ab *y es primo relativo con* a, r *divide a* b. (*Indicación*: *si* r *es* primo relativo con a, existen dos enteros k y l tales que

$$kr + la = 1.$$

Multiplíquense los dos miembros de esta igualdad por b.) Este teorema incluye como caso particular el lema de la página 54, ya que un número primo p es primo relativo con un entero a cuando, y solamente entonces, p no divide a a.

Para todo entero positivo n, designemos con $\varphi(n)$ *el número de enteros primos con* n *y menores que* n. Esta función, introducida por Euler, es una función de gran importancia en teoría de números. Los valores de $\varphi(n)$ para los primeros valores de n pueden calcularse fácilmente:

$\varphi(1) = 1$	puesto que	1	es primo relativo		con 1,
$\varphi(2) = 1$	»	1	»	»	con 2,
$\varphi(3) = 2$	»	1 y 2	son primos relativos		con 3,
$\varphi(4) = 2$	»	1 y 3	»	»	con 4,
$\varphi(5) = 4$	»	1, 2, 3, 4	»	»	con 5,
$\varphi(6) = 2$	»	1, 5	»	»	con 6,
$\varphi(7) = 6$	»	1, 2, 3, 4, 5, 6	son primos relativos		con 7,
$\varphi(8) = 4$	»	1, 3, 5, 7	»	»	con 8,
$\varphi(9) = 6$	»	1, 2, 4, 5, 7, 8	»	»	con 9,
$\varphi(10) = 4$	»	1, 3, 7, 9	»	»	con 10.
etc.					

Se observa que $\varphi(p) = p - 1$ si p es primo, pues un número primo p no tiene más divisores que él mismo y la unidad, y, por tanto, es primo relativo con todos los enteros 1, 2, 3, ..., $p - 1$. Si n es compuesto, con la descomposición en factores primos

$$n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r},$$

donde las p representan números primos distintos, cada uno elevado a una cierta potencia, se tiene

$$\varphi(n) = n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right)\cdots\left(1 - \frac{1}{p_r}\right)$$

Por ej., puesto que $12 = 2^2 \cdot 3$,

$$\varphi(12) = 12(1 - 1/2)(1 - 1/3) = 12(1/2)(2/3) = 4,$$

como debía resultar, previo el cómputo de primos relativos con 12 y menores que él. La demostración general es completamente elemental, pero no la daremos aquí.

***Ejercicio:** Basándose en la función φ de Euler, generalícese el teorema de Fermat de la página 44. El teorema general se enuncia así: *Si* n *es un entero cualquiera, y* a *es primo relativo con* n, *se tiene*

$$a^{\varphi(n)} \equiv 1 \qquad (\text{mód } n).$$

4. Fracciones continuas. Ecuaciones diofánticas.—El algoritmo de Euclides para el cálculo del máximo común divisor de dos enteros conduce de modo natural a un importante método para representar el cociente de dos enteros mediante una fracción compuesta; p. ej., aplicando a los números 840 y 611 el algoritmo de Euclides resultan las igualdades

$$840 = 1 \cdot 611 + 229, \qquad 611 = 2 \cdot 229 + 153,$$
$$229 = 1 \cdot 153 + 76, \qquad 153 = 2 \cdot 76 + 1,$$

las cuales demuestran, incidentalmente, que (840, 611) = 1. De estas ecuaciones se pueden deducir las siguientes expresiones:

$$\frac{840}{611} = 1 + \frac{229}{611} = 1 + \frac{1}{611/229},$$

$$\frac{611}{229} = 2 + \frac{153}{229} = 2 + \frac{1}{229/153},$$

$$\frac{229}{153} = 1 + \frac{76}{153} = 1 + \frac{1}{153/76},$$

$$\frac{153}{76} = 2 + \frac{1}{76}.$$

Por combinación de estas igualdades se obtiene el desarrollo del número racional $\frac{840}{611}$ en la forma

$$\frac{840}{611} = 1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{76}}}}$$

Una expresión de la forma

$$a = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots + \cfrac{1}{a_n}}}, \qquad [7]$$

donde las a representan enteros positivos, se llama *fracción continua*. El algoritmo de Euclides da un método para representar todo número fraccionario en esta forma.

Ejercicio: Hállense los desarrollos en fracción continua de

$$\frac{2}{5}, \frac{43}{30}, \frac{169}{70}$$

*Las fracciones continuas son de gran importancia en la rama de la aritmética superior conocida con el nombre de análisis diofántico. Una *ecuación diofántica* es una ecuación algebraica, con una o más incógnitas, de coeficientes *enteros* y de la que interesan únicamente las soluciones *enteras*. Tales ecuaciones pueden carecer de soluciones o tener un número finito o infinito de ellas. El caso más sencillo es el de la ecuación diofántica *lineal* con dos incógnitas,

$$ax + by = c, \qquad [8]$$

donde a, b y c son enteros dados, y se buscan soluciones x, y enteras. La solución completa de una ecuación de esta forma puede hallarse por el algoritmo de Euclides.

Para comenzar, hallemos $d = (a, b)$ por el algoritmo de Euclides; se tendrá entonces, para k y l convenientes,

$$ak + bl = d, \qquad [9]$$

de donde resulta que la ecuación [8] admitirá la solución particular $x = k$, $y = l$ en el caso en que sea $c = d$. En general, si c es un múltiplo de d,

$$c = d \cdot q,$$

se obtendría de [9]

$$a(kq) + b(lq) = dq = c,$$

de modo que [8] tiene la solución particular $x = x^* = kq$, $y = y^* = lq$. Recíprocamente: si [8] tiene una solución x, y para un c dado, c debe ser múltiplo de $d = (a, b)$, puesto que q, por dividir a a y b, debe dividir a c. Hemos probado así que, para que la ecuación [8] admita soluciones, es necesario y suficiente que c sea múltiplo de (a, b).

Para determinar otras posibles soluciones de [8] observemos que si $x = x'$, $y = y'$ es otra solución cualquiera, distinta de la $x = x^*$, $y = y^*$ obtenida por el algoritmo de Euclides, las diferencias $x = x' - x^*$, $y = y' - y^*$ forman una solución de la ecuación «homogénea»

$$ax + by = 0. \qquad [10]$$

Pues si es

$$ax' + by' = c \qquad \text{y} \qquad ax^* + by^* = c,$$

restando miembro a miembro estas dos ecuaciones se obtiene

$$a(x' - x^*) + b(y' - y^*) = 0.$$

Ahora bien: la solución general de la ecuación [10] es $x = rb/(a, b)$, $y = -ra/(a, b)$, donde r es un entero cualquiera. (Dejamos al lector la demostración como ejercicio.

Indicación: Divídase por (a, b) y utilícese el ejercicio de la pág. 56.) Resulta entonces de modo inmediato que

$$x = x^* + rb/(a, b), \qquad y = y^* - ra/(a, b).$$

En resumen: La ecuación diofántica lineal $ax + by = c$, donde a, b y c son enteros, tiene soluciones cuando, y únicamente entonces, c es múltiplo de (a, b). En este caso, se puede hallar una solución por el algoritmo de Euclides, y la solución general es de la forma

$$x = x^* + rb/(a, b), \qquad y = y^* - ra/(a, b),$$

donde r es un entero cualquiera.

Ejemplos: La ecuación $3x + 6y = 22$ no admite soluciones enteras, puesto que $(3, 6) = 3$ no divide a 22.

La ecuación $7x + 11y = 13$ tiene la solución particular $x = -39$, $y = 26$, obtenida en la forma siguiente:

$$11 = 1 \cdot 7 + 4, \qquad 7 = 1 \cdot 4 + 3, \qquad 4 = 1 \cdot 3 + 1, \qquad (7, 11) = 1.$$
$$1 = 4 - 3 = 4 - (7 - 4) = 2 \cdot 4 - 7 = 2(11 - 7) - 7 = 2 \cdot 11 - 3 \cdot 7.$$

De donde

$$7 \cdot (-3) + 11(2) = 1,$$
$$7 \cdot (-39) + 11(26) = 13.$$

Las demás soluciones vienen dadas por las fórmulas

$$x = -39 + 11r, \qquad y = 26 - 7r,$$

siendo r un entero arbitrario.

Ejercicio: Resuélvanse las ecuaciones diofánticas: *a)* $3x - 4y = 29$; *b)* $11x + 12y = 58$; *c)* $153x - 34y = 51$.

CAPÍTULO II

SISTEMAS DE NÚMEROS

Introducción.—Debemos extender suficientemente el concepto inicial de número, como número natural, hasta crear un instrumento capaz de satisfacer las necesidades de la práctica y de la teoría. En una larga y, a veces, titubeante evolución histórica fueron gradualmente aceptados el cero, los enteros negativos y las fracciones, en el mismo plano que los enteros positivos, y hoy día las reglas operativas con estos números son del dominio de todo estudiante de bachillerato. Pero para alcanzar completa libertad en las operaciones algebraicas debemos ir más allá, hasta incluir en el concepto de número las cantidades irracionales y complejas. Aunque estas extensiones del concepto de número natural han sido utilizadas en matemática durante varios siglos y, por otra parte, constituyen la base de toda la matemática moderna, hasta tiempos relativamente recientes no fueron establecidas sobre una base lógica sólida. En el presente capítulo haremos una exposición del modo como dicha base fué alcanzada.

I. LOS NÚMEROS RACIONALES

1. Los números racionales como resultado de mediciones.—Los números enteros son abstracciones del proceso de contar colecciones finitas de objetos. Pero en la vida diaria no es suficiente poder *contar objetos* individuales, es preciso también *medir cantidades* tales como longitudes, áreas, pesos y tiempo. Si se quiere operar sin obstáculos con las medidas de estas cantidades, que son susceptibles de subdivisiones arbitrariamente pequeñas, es necesario extender el campo de la aritmética más allá de los números enteros. El primer paso será *el de reducir el problema de la medida al de contar*. Comenzaremos por elegir, de modo completamente arbitrario, una *unidad de medida* —metro, pie, gramo, libra, segundo, etc.—a la que asignaremos la medida 1. Luego, contaremos el número de esas unidades contenidas en la cantidad que deseamos medir; p. ej., una cierta masa de plomo pesa exactamente 54 Kg. Sin embargo, el proceso de contar no es suficiente en general, ya que la cantidad dada puede no ser exactamente medible mediante múltiplos enteros de la unidad elegida. Las más de las veces podremos decir únicamente que dicha cantidad está

comprendida entre dos múltiplos consecutivos de la unidad; p. ej., entre 53 Kg y 54 Kg. Cuando esto ocurra, avanzaremos un paso introduciendo nuevas subunidades, obtenidas por subdivisión de la unidad inicial en un cierto número n de partes iguales. En el lenguaje ordinario, estas nuevas subunidades pueden tener nombres especiales; p. ej., el pie se divide en 12 pulgadas; el metro, en 100 centímetros; la libra, en 16 onzas; la hora, en 60 minutos; el minuto, en 60 segundos, etc. Sin embargo, en el simbolismo de las matemáticas, una subunidad obtenida dividiendo la unidad inicial en n partes iguales se designa con el símbolo $1/n$; y si una cantidad contiene exactamente m de estas subunidades, su medida se denota con el símbolo m/n. Este símbolo se llama *fracción* o *razón* (a veces se escribe $m : n$). El paso siguiente, verdaderamente decisivo, sólo se dió de modo consciente después de varios siglos de tentativas. El resultado fué que el símbolo m/n quedó desposeído de referencias concretas a procesos de medidas y a las cantidades medidas, y fué considerado simplemente como un *número*, un ente en sí mismo, en el mismo plano que los números naturales. Cuando m y n son números naturales, el símbolo m/n se llama *número racional*.

El uso de la palabra número (inicialmente reservada para los números naturales) para estos nuevos símbolos está justificado por el hecho de que la adición y la multiplicación de estos entes obedecen a las mismas leyes que rigen dichas operaciones con los números naturales. Para probar esto, debemos definir previamente la adición, la multiplicación y la igualdad de números racionales. Como es bien sabido, estas definiciones son:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}, \qquad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd} \tag{1}$$

$$\frac{a}{a} = 1, \qquad \frac{a}{b} = \frac{c}{d} \text{ si } ad = bc,$$

para enteros cualesquiera a, b, c, d; p. ej.:

$$\frac{2}{3} + \frac{4}{5} = \frac{2 \cdot 5 + 3 \cdot 4}{3 \cdot 5} = \frac{10 + 12}{15} = \frac{22}{15}, \qquad \frac{2}{3} \cdot \frac{4}{5} = \frac{2 \cdot 4}{3 \cdot 5} = \frac{8}{15}.$$

$$\frac{3}{3} = 1, \qquad \frac{8}{12} = \frac{6}{9} = \frac{2}{3}$$

Estas definiciones se nos presentan forzosamente si deseamos que los números racionales sean apropiados para medir longitudes, áreas, etc. Pero hablando en sentido estricto, estas reglas de adición, multiplicación e igualdad de nuestros símbolos quedan establecidas por su pro-

pia definición y no aparecen impuestas por otras necesidades que las de ser no contradictorias y resultar útiles para las aplicaciones. A partir de las definiciones [1] se puede probar que *las leyes fundamentales de la aritmética de los números naturales continúan siendo válidas en el dominio de los números racionales*:

$$
\begin{aligned}
p + q &= q + p && \text{(ley conmutativa de la adición),} \\
p + (q + r) &= (p + q) + r && \text{(ley asociativa de la adición),} \\
pq &= qp && \text{(ley conmutativa de la multiplicación),} \qquad [2] \\
p(qr) &= (pq)r && \text{(ley asociativa de la multiplicación),} \\
p(q + r) &= pq + pr && \text{(ley distributiva).}
\end{aligned}
$$

P. ej., la prueba de la ley conmutativa de la adición de fracciones resulta de las ecuaciones

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} = \frac{cb + da}{db} = \frac{c}{d} + \frac{a}{b},$$

en las cuales el primero y el último signo de igualdad corresponden a la definición [1] de la adición, mientras que el del centro es una consecuencia de las leyes conmutativas de la adición y de la multiplicación de números naturales. El lector puede comprobar las otras cuatro leyes de manera análoga.

Para la efectiva comprensión de estos hechos se debe insistir una vez más en que los números racionales son creación nuestra, y que las reglas [1] dependen de nuestra voluntad. Podríamos haber definido caprichosamente la adición por la fórmula $\frac{a}{b} + \frac{c}{d} = \frac{a + c}{b + d}$, la cual habría dado en particular $^1/_2 + ^1/_2 = ^2/_4$, lo que sería absurdo aplicado a la medición de cantidades. Reglas de tal tipo, aunque permisibles lógicamente, harían de la aritmética de nuestros nuevos símbolos un juego carente de sentido. El juego libre del intelecto debe estar guiado aquí por la necesidad de crear un instrumento capaz de ser utilizado para la medida.

2. Necesidad intrínseca de la introducción de los números racionales. Principio de generalización.—Junto a las razones *prácticas* que indujeron a la introducción de los números racionales, existen otras de carácter intrínseco y en cierto modo más apremiantes, que vamos a discutir independientemente de los argumentos anteriores. Estas razones son de carácter aritmético y típicas de una tendencia dominante en el proceso matemático.

En la aritmética ordinaria de los números naturales se pueden efectuar siempre las dos operaciones fundamentales: adición y multiplicación. En cambio, las *operaciones inversas* no son siempre posi-

bles. La diferencia $b - a$ de dos enteros a, b es el entero c tal que $a + c = b$; es decir, es la solución de la ecuación $a + x = b$. Pero en el dominio de los números naturales el símbolo $b - a$ posee significación únicamente cuando es $b > a$, ya que únicamente entonces tiene la ecuación $a + x = b$ una solución que sea un número natural. Un gran paso para suprimir esta restricción se dió cuando se introdujo el símbolo 0 mediante la relación $a - a = 0$. De mayor importancia aún fué la introducción de los símbolos -1, -2, -3, ..., junto con la definición

$$b - a = -(a - b)$$

para el caso $b < a$, que permitió la sustracción, sin restricciones, *en el dominio de los enteros positivos y negativos.* Para incluir los nuevos símbolos -1, -2, -3, ..., en una aritmética más amplia, que comprenda tanto los enteros positivos como los negativos, debemos, naturalmente, *definir las operaciones* con ellos de tal manera que *las reglas de las operaciones aritméticas con los números naturales se conserven para el nuevo dominio*; p. ej., la regla

$$(-1)(-1) = 1, \qquad [3]$$

que servirá para regir la multiplicación de los enteros negativos, es una consecuencia del deseo de conservar la ley distributiva $a(b + c) = = ab + ac$. Puesto que si hubiéramos convenido, p. ej., en que fuera $(-1)\ (-1) = -1$, poniendo $a = -1$, $b = 1$, $c = -1$ resultaría $-1\ (1 - 1) = -1 - 1 = -2$, mientras que por otro lado se tendría $-1(1 - 1) = -1 \cdot 0 = 0$. Fué necesario mucho tiempo para que los matemáticos comprendieran que la «regla de los signos» [3], junto con todas las demás definiciones que se refieren a los enteros negativos y a las fracciones, no podían ser «demostradas». Todas eran *creaciones* hechas con objeto de alcanzar libertad en las operaciones, conservando siempre las leyes fundamentales de la aritmética. Lo que *puede*—y debe—probarse es únicamente el hecho de que con tales definiciones las leyes conmutativa, asociativa y distributiva de la aritmética se conservan. Aun el gran matemático Euler dió un argumento poco convincente para mostrar que $(-1)\ (-1)$ «debe» ser igual a $+1$. Decía: dicho producto debe ser $+1$ ó -1, pero no puede ser -1, puesto que $-1 = (+1)\ (-1)$.

Del mismo modo que la introducción de los enteros negativos y del cero despejó el camino para la sustracción sin restricciones, la introducción de los números fraccionarios suprime análogos obstáculos para la división. El cociente $x = b/a$ de dos enteros a y b, definido por la ecuación

$$ax = b, \qquad [4]$$

existe, *como entero*, únicamente cuando a es un divisor de b. Si no es ése el caso, como, p. ej., si es $a = 2$, $b = 3$, introducimos simplemente el símbolo b/a, al que llamamos fracción, para el que establecemos la regla $a(b/a)$, $= a$, de modo que b/a es, «por definición», una solución de [4]. La introducción de las fracciones como nuevos números hace posible la división sin restricciones, *excepto la división por cero*, la cual será *excluída en todos los casos.*

Expresiones del tipo $1/0$, $3/0$, $0/0$, etc., serán siempre símbolos sin significado para nosotros, puesto que si se admitiera la división por cero, podríamos deducir de ecuaciones correctas, como $0 \cdot 1 = 0 \cdot 2$, la consecuencia absurda $1 = 2$. Sin embargo, a veces puede ser útil designar expresiones del tipo $3/0$ por el símbolo ∞ (léase «infinito»), *siempre que no se pretenda operar con el símbolo* ∞ *como si estuviera sujeto a las leyes ordinarias del cálculo numérico.*

La significación aritmética propia del sistema de *todos los números racionales*—enteros y fraccionarios, negativos y positivos—resulta ahora clara. Para el dominio de los números así extendido no sólo valen las leyes formales asociativa, conmutativa y distributiva, sino que ahora también las ecuaciones $a + x = b$ y $ax = b$ tienen las soluciones $x = b - a$ y $x = b/a$, sin restricción alguna, con tal que para la última se tenga $a \neq 0$. En otros términos, en el dominio de los números racionales las llamadas *operaciones racionales*—adición, sustracción, multiplicación y división—son posibles sin restricción y los resultados pertenecen siempre a aquel dominio de números. Un dominio de números *cerrado* respecto de dichas operaciones se llama un *cuerpo*. Daremos otros ejemplos de cuerpos más adelante, en este capítulo y en el siguiente.

Extender un dominio por la introducción de nuevos símbolos, de tal modo que las leyes que valen en el primero continúen rigiendo en el segundo, es uno de los aspectos del proceso de *generalización* característico de la matemática. La generalización del concepto de número natural al de número racional satisface, por una parte, la necesidad teórica de suprimir las restricciones a la sustracción y a la división, y cumple, por otra, la necesidad práctica de tener números para representar los resultados de mediciones. Del hecho de que los números racionales satisfagan esa doble necesidad resulta verdaderamente su gran importancia. Como hemos visto, esta extensión del concepto de número ha sido posible por la creación de nuevos números en la forma de símbolos abstractos tales como 0, -2, y $^3/_4$. Hoy, acostumbrados como estamos a tratarlos como cosa corriente, resulta difícil creer que hasta el siglo XVII no fueron admitidos con los mismos derechos

que los enteros positivos y que, aunque usados cuando se hacían necesarios, no era sin ciertas dudas y prevenciones. A la natural tendencia humana a apoyarse en lo *concreto*, y como tales aparecían los números naturales, se debe la lentitud con que se dió este paso inevitable. Únicamente en el dominio de lo abstracto puede ser creado un sistema satisfactorio de aritmética.

3. Interpretación geométrica de los números racionales.—La construcción que sigue dará una interpretación geométrica intuitiva del sistema de los números racionales.

Tomemos sobre una recta, «recta numérica», un segmento de 0 a 1 (Fig. 8). Elijamos dicho segmento como unidad de longitudes, unidad que puede ser tomada arbitrariamente. Los enteros positivos y negativos serán representados por puntos equidistantes sobre la recta numérica, los positivos a la derecha del 0 y los negativos a la izquierda.

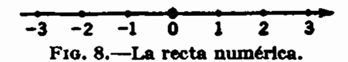


FIG. 8.—La recta numérica.

Para representar las fracciones de denominador n, dividimos cada uno de los segmentos de longitud unidad en n partes iguales; los puntos de subdivisión representan las fracciones con denominador n. Si efectuamos esa construcción para todo entero n, todos los números racionales vendrán representados por puntos de la recta numérica. Llamaremos a los puntos así obtenidos *puntos racionales*, y usaremos las expresiones «número racional» y «punto racional» como equivalentes.

En el capítulo primero, I, se definió la relación $A < B$ para números naturales. Esta relación tiene una interpretación en la recta numérica, que resulta del hecho de que si un número natural A es menor que otro B, el punto A está a la izquierda del punto B. Esta relación geométrica tiene significado para *todos* los puntos racionales, por lo que parece natural intentar extender la relación aritmética a los números racionales, de modo que corresponda al orden geométrico de los puntos racionales. Se llega a ese resultado con la definición siguiente: Diremos que el número racional A es *menor que* el número racional B ($A < B$), y B se dice *mayor que* A ($B > A$) cuando $B - A$ es positivo. De la definición se sigue que, si es $A < B$, los puntos (números) *entre* A y B son aquellos que satisfacen simultáneamente las condiciones de ser $> A$ y $< B$. Todo par de puntos distintos, junto con los puntos comprendidos entre ellos, forman el *segmento* o *intervalo* $[A, B]$.

La distancia de un punto A al origen, considerada como positiva, se llama *valor absoluto* de A y se indica con el símbolo

$$|A|$$

Es decir; si es $A \geqslant 0$, se tiene $|A| = A$; y si es $A \leqslant 0$, se tiene $|A| = -A$. Es fácil ver que si A y B tienen el mismo signo, vale la igualdad $|A + B| = |A| + |B|$; mientras que si A y B tienen signos distintos, es $|A + B| < |A| + |B|$. Por tanto, combinando estos dos resultados, se tendrá en todo caso

$$|A + B| \leqslant |A| + |B|,$$

cualesquiera que sean los signos de A y B.

Para la representación que consideramos se tiene la siguiente proposición de fundamental importancia: *Los puntos racionales forman un conjunto denso en la recta.* Con esto se quiere decir que interiores a todo intervalo, por pequeño que sea, hay siempre puntos racionales. Basta tomar el denominador n suficientemente grande, de modo que el intervalo $[0,1/n]$ sea más pequeño que el intervalo $[A, B]$ en cuestión, para que al menos una de las fracciones m/n sea interior a él. No existen, por tanto, intervalos, por pequeños que sean, vacíos de puntos racionales. Resulta también que en todo intervalo debe haber infinitos puntos racionales; puesto que si hubiera solamente un número finito de ellos, el intervalo determinado por dos consecutivos no podría contener puntos racionales, en oposición a lo que acabamos de probar.

II. SEGMENTOS INCONMENSURABLES, NÚMEROS IRRACIONALES Y CONCEPTO DE LÍMITE

1. Introducción.—Cuando se comparan las longitudes de dos segmentos rectilíneos a y b, puede ocurrir que a esté contenido un número r exacto de veces en b. En este caso podemos expresar la medida del segmento b tomando como unidad a y diciendo que la longitud de b es r veces la de a. Pero puede ocurrir que, mientras que ningún múltiplo entero de a sea igual a b, se pueda dividir a en un cierto número de partes iguales, p. ej., n, cada una de longitud a/n y tales que un múltiplo entero m del segmento a/n sea igual a b:

$$b = \frac{m}{n} a. \qquad [1]$$

Cuando se tiene una igualdad de la forma [1] diremos que los dos segmentos a y b son *conmensurables,* dado que tienen una medida común: el segmento a/n está contenido n veces en a y m veces en b. El conjunto de todos los segmentos conmensurables con a estará constituído por aquellos segmentos cuya longitud puede ser expresada en

la forma [1], para enteros m y n convenientes ($n \neq 0$). Si tomamos a como segmento unidad [0,1], en la figura 9, los segmentos conmensu-

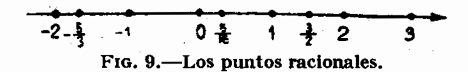


FIG. 9.—Los puntos racionales.

rables con el segmento unidad corresponderán a todos los puntos racionales m/n sobre la recta numérica. Para todas las cuestiones prácticas relacionadas con la medida, los números racionales son suficientes. Incluso desde un punto de vista teórico, como los puntos racionales cubren la recta densamente, podría pensarse que todos los puntos de la recta fueran racionales. Si esta sospecha fuese cierta, todos los segmentos serían conmensurables con la unidad. Uno de los descubrimientos más sorprendentes de los primeros matemáticos griegos (la escuela pitagórica) fué el de que las cosas no sucedían de modo tan simple. Existen *segmentos inconmensurables* y, si suponemos que a todo segmento corresponde un número, como medida de su longitud con un segmento unidad, existen también *números irracionales*. Este descubrimiento fué un acontecimiento científico de la máxima importancia. Posiblemente señala el origen de lo que puede considerarse como contribución específica de los griegos a los procesos rigurosos de las matemáticas. Es evidente que este hecho afectó profundamente la matemática y la filosofía desde la época griega hasta nuestros días.

La teoría de los inconmensurables de Eudoxio, presentada en forma geométrica en los *Elementos* de Euclides, es una obra maestra de la matemática griega, frecuentemente omitida en las desfiguradas versiones didácticas de los *Elementos*. Dicha teoría no fué apreciada en su justo valor hasta el siglo pasado, después que Dedekind, Cantor y Weierstrass construyeron una teoría rigurosa de los números irracionales. Expondremos aquí la teoría de los inconmensurables en la forma aritmética moderna.

Antes de nada probaremos que *la diagonal del cuadrado es inconmensurable con su lado*. Supongamos que se ha tomado como unidad de longitud el lado del cuadrado y que la diagonal tiene una longitud x. Entonces, por el teorema de Pitágoras, se tendrá

$$x^2 = 1^2 + 1^2 = 2.$$

(Designamos x con el símbolo $\sqrt{2}$.) Si x fuese conmensurable con 1, existirían dos enteros p y q tales que $x = p/q$, y

$$p^2 = 2q^2. \qquad [2]$$

Supongamos p/q irreducible, ya que puede suprimirse todo factor común al numerador y denominador. Puesto que 2 aparece como factor en el segundo miembro, p^2 tiene que ser un número par, de donde resulta que p debe ser par, pues el cuadrado de todo número impar es impar. Podemos, por tanto, escribir $p = 2r$. Al sustituir este valor en [2] se tiene

$$4r^2 = 2q^2, \quad \text{o lo que es lo mismo,} \quad 2r^2 = q^2.$$

Por ser 2 un factor del primer miembro de la última igualdad, q^2 es par y, en consecuencia, también lo es q. Resultan así p y q divisibles por 2, lo que contradice la hipótesis de que p y q no tenían factores comunes; en consecuencia, no se puede tener la igualdad [2]; es decir, x no puede ser un número racional.

Nuestro resultado puede ser expresado por la proposición: no existe ningún número racional igual a $\sqrt{2}$.

Del argumento precedente resulta que, mediante una construcción geométrica muy sencilla, se puede obtener un segmento inconmensurable con el segmento unidad. Si con un compás llevamos dicho segmento sobre la recta numérica, en la forma indicada en la figura 10, el punto obtenido no puede coincidir con ninguno de los puntos racionales: *El sistema de los números racionales*, aunque denso en toda ella, *no cubre toda la recta numérica*. A una mente ingenua debe resultarle ciertamente extraño y paradójico que el conjunto denso de los puntos racionales no llena la recta completa. Nada en nuestra «intuición» puede ayudarnos a «ver» los puntos irracionales como distintos de los racionales. No es de extrañar que el descubrimiento de los inconmensurables produjera gran impresión en los filósofos y matemáticos griegos, y que aún hoy le ofrezca, a quien medite sobre la cuestión, cierta perplejidad.

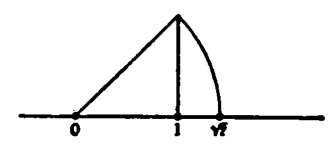


FIG. 10.—Construcción de $\sqrt{2}$.

Es fácil construir tantos segmentos inconmensurables con la unidad como se desee. Los extremos de tales segmentos llevados a partir del 0 de la recta numérica son los llamados *puntos irracionales*. Ahora bien: el principio que sirvió para introducir las fracciones fué el de *medir las longitudes con números*, y deseamos en lo que sigue conservar este principio y tratar de acuerdo con él los segmentos inconmensurables con la unidad. Si queremos que exista una *correspondencia mutua entre números*, de una parte, *y puntos de la recta*, de otra, es preciso introducir los *números irracionales*.

Resumiendo en lo posible la situación, se puede decir que un número irracional representa la longitud de un segmento inconmensurable con la unidad. En las secciones que siguen precisaremos esta definición un poco vaga y completamente geométrica, hasta llegar a otra más satisfactoria desde el punto de vista del rigor lógico. Nuestras primeras consideraciones a este propósito partirán de las propiedades de las fracciones decimales.

Ejercicios:

1. Demuéstrese que $\sqrt[3]{2}$, $\sqrt{3}$, $\sqrt{5}$, $\sqrt[3]{3}$ no son racionales. (*Indicación:* Utilícese el lema de la página 54.)
2. Pruébese que $\sqrt{2} + \sqrt{3}$ y $\sqrt{2} + \sqrt[3]{2}$ no son racionales. (*Indicación:* Si, p. ej., el primero de estos números fuera igual a un número racional r, poniendo $\sqrt{3} = r - \sqrt{2}$ y elevando al cuadrado, $\sqrt{2}$ resultaría racional.)
3. Demuéstrese que $\sqrt{2} + \sqrt{3} + \sqrt{5}$ es irracional. Inténtese construir ejemplos análogos y más generales.

2. Fracciones decimales. Decimales de infinitas cifras.—Para cubrir la recta numérica con un conjunto de puntos denso en toda ella no son necesarios *todos* los números racionales; bastan, por ejemplo, los números obtenidos por subdivisión de cada intervalo unidad en 10, luego en 100, 1000, etc., segmentos iguales. Los puntos así obtenidos corresponden a las «fracciones decimales»; p. ej., el punto $0{,}12 = 1/10 + 2/100$ corresponde al punto situado en el primer intervalo unidad, en el segundo subintervalo de longitud 10^{-1}, y en el origen del tercer «sub-sub-» intervalo de longitud 10^{-2}. (a^{-n} significa $1/a^n$.) Si una *fracción decimal* contiene n cifras después de la coma, tiene la forma

$$f = z + a_1 10^{-1} + a_2 10^{-2} + a_3 10^{-3} + \ldots + a_n 10^{-n},$$

donde z es un entero y las a son cifras—0, 1, 2, ..., 9—que indican las décimas, centésimas, y así sucesivamente. El número f se representa en el sistema decimal en la forma abreviada $z{,}a_1a_2a_3\ldots a_n$. Se ve inmediatamente que estas fracciones pueden escribirse en forma de fracción p/q, siendo $q = 10^n$; p. ej., $f = 1{,}314 = 1 + 3/10 + 1/100 + 4/1000 = 1314/1000$. Si p y q tienen un divisor común, la fracción podrá reducirse a otra cuyo denominador será divisor de 10^n. Por otra parte, ninguna fracción irreducible cuyo denominador no sea divisor de alguna potencia de 10 puede venir representada por una fracción decimal; p. ej., $1/5 = 2/10 = 0{,}2$ y $1/250 = 4/1000 = 0{,}004$; en cambio, $1/3$ no puede ser escrita como número decimal de n cifras, por grande que sea n, ya que una igualdad de la forma

$$1/3 = b/10^n$$

conduciría a

$$10^n = 3b,$$

lo que es absurdo, ya que 3 no es factor de ninguna potencia de 10.

Tomemos sobre la recta numérica un punto P que no corresponda a ninguna fracción decimal; p. ej., el punto racional 1/3 o el punto irracional $\sqrt{2}$. Entonces, por el proceso de subdivisión del intervalo unidad en diez partes iguales, y luego en cien y así sucesivamente, el punto P no será nunca origen de ninguno de los subintervalos parciales. Sin embargo, P puede ser incluído dentro de intervalos cada vez más pequeños de la subdivisión decimal, con el grado de aproximación que se desee. Este proceso de aproximación puede ser descrito en la forma siguiente: supongamos que P está en el primer intervalo unidad; subdividiendo este intervalo en 10 partes iguales, cada una de longitud 10^{-1}, supongamos que P está en el tercero de tales intervalos. Diremos entonces que P está entre las fracciones decimales 0,2 y 0,3. Subdividimos entonces el intervalo de 0,2 a 0,3 en 10 partes iguales, cada una de longitud 10^{-2}, y supongamos que P está en el cuarto de tales intervalos. Subdividamos éste a su vez y supóngase que P está en el primero de estos intervalos de longitud 10^{-3}. Podremos decir entonces que P está entre 0,230 y 0,231. Este proceso puede continuarse indefinidamente, dando lugar a una sucesión ilimitada de cifras, $a_1, a_2, a_3, \ldots, a_n, \ldots$, con la propiedad siguiente: para todo valor del entero n, el punto P está incluído en el intervalo I_n, cuyo origen es la fracción decimal $0{,}a_1a_2a_3 \ldots a_{n-1}a_n$ y cuyo extremo es $0{,}a_1a_2a_3 \ldots a_{n-1}(a_n + 1)$, siendo 10^{-n} la longitud de I_n. Si elegimos la sucesión $n = 1, 2, 3, 4, \ldots$, vemos que cada uno de los intervalos $I_1, I_2, I_3, \ldots$, está contenido en el precedente, mientras que sus longitudes 10^{-1}, 10^{-2}, 10^{-3}, ..., tienden a cero. Diremos que P está contenido en una sucesión de *intervalos decimales encajados;* p. ej., si P es el punto racional 1/3, todas las cifras $a_1, a_2, a_3, \ldots$ son iguales a 3, y P está contenido en todo intervalo I_n cuyos extremos sean 0,333...33 y 0,333...34; es decir, 1/3 es mayor que 0,333...33 y menor que 0,333...34, donde el número de cifras puede ser arbitrariamente grande. Expresaremos este hecho diciendo que el número decimal de n cifras 0,333 ... 33 «tiende hacia 1/3» al crecer n, y escribiremos

$$1/3 = 0{,}333\ldots;$$

los puntos indican que la fracción decimal debe extenderse «indefinidamente».

El punto irracional $\sqrt{2}$ antes definido conduce también a una frac-

ción decimal que se extiende indefinidamente; sin embargo, en este caso, la ley que determina los valores de las cifras de la sucesión no es sencilla. En efecto, no se conoce ninguna fórmula explícita que determine las cifras de la sucesión, aunque se pueden calcular tantas como se desee:

$$1^2 = 1 < 2 < 2^2 = 4$$
$$(1{,}4)^2 = 1{,}96 < 2 < (1{,}5)^2 = 2{,}25$$
$$(1{,}41)^2 = 1{,}9881 < 2 < (1{,}42)^2 = 2{,}0264$$
$$(1{,}414)^2 = 1{,}999396 < 2 < (1{,}415)^2 = 2{,}002225$$
$$(1{,}4142)^2 = 1{,}99996164 < 2 < (1{,}4143)^2 = 2{,}00024449, \text{ etc.}$$

Como definición general diremos que un punto P que no esté representado por una fracción decimal con un número finito n de cifras, está representado por la *fracción decimal infinita*, $z{,}a_1a_2a_3 \ldots$, si para cualquier valor de n el punto P está situado en el intervalo de longitud 10^{-n} con origen en el punto $z{,}a_1a_2a_3 \ldots a_n$.

De este modo se establece una correspondencia entre los puntos de la recta numérica y todas las fracciones decimales *finitas* e *infinitas*. Esto sugiere la siguiente definición: un «número» es una fracción decimal *finita* o *infinita*. Los decimales infinitos que no representan números racionales serán llamados *números irracionales*.

Hasta mediados del siglo pasado, las consideraciones precedentes eran aceptadas como una exposición satisfactoria del sistema de los números racionales e irracionales, sistema designado con el nombre de *continuo numérico*. El avance enorme de las matemáticas a partir del siglo XVII, en particular el desarrollo de la geometría analítica y del cálculo diferencial e integral, se hace sin riesgo con este concepto de sistema numérico como base. Pero durante el período de examen crítico de los principios y de consolidación de los resultados, fué abriéndose paso la idea de que el concepto de número irracional requería un análisis más preciso. Como preliminar a nuestra exposición de la teoría moderna del continuo numérico, discutiremos en forma más o menos intuitiva el concepto básico de *límite*.

Ejercicio: Calcúlese $\sqrt[3]{2}$ y $\sqrt[3]{5}$ con una aproximación de 10^{-2}.

3. Límites. Progresiones geométricas indefinidas.—Como vimos en la sección precedente, ocurre a veces que un cierto número racional s viene aproximado por una sucesión de otros números racionales s_n, en la cual el índice n toma sucesivamente los valores 1, 2, 3, ... ; p. ej., si es $s = 1/3$, la sucesión de números racionales $s_1 = 0{,}3$, $s_2 = 0{,}33$, $s_3 = 0{,}333$, etc., tiende hacia s. Para tener otro ejemplo, dividamos el

intervalo unidad en dos mitades, la segunda mitad en otras dos partes iguales, y así sucesivamente, hasta que los dos menores intervalos obtenidos sean de longitud 2^{-n}, donde n es arbitrariamente grande; es decir, $n = 100$, $n = 100\,000$, o el número que queramos. Sumando entonces todos los intervalos, excepto el último, obtenemos una longitud total igual a

$$s_n = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots + \frac{1}{2^n} \qquad [3]$$

Vemos que s_n difiere de 1 en $(^1/_2)^n$, y que esta diferencia llega a ser arbitrariamente pequeña, o «tiende a cero» cuando n crece indefinidamente. Carece de sentido decir que la diferencia *es* cero cuando n *es* infinito. El infinito aparece únicamente en el *proceso* y no como una *cantidad* efectiva. Podemos describir el comportamiento de s_n diciendo que *la suma* s_n *se aproxima a 1 cuando* n *tiende a infinito*, y escribir

$$1 = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \cdots, \qquad [4]$$

donde en el segundo miembro se tiene una *serie indefinida* o *infinita*. Esta «ecuación» no significa que debamos sumar efectivamente infinitos sumandos; es sólo una expresión abreviada para el hecho de que 1 es el límite de la suma finita s_n cuando n *tiende a* infinito (no que *es* infinito). Así, la ecuación [4], con su símbolo incompleto «$+ \ldots$», es sólo una manera breve de escribir la proposición precisa:

1 = al límite, cuando n tiende hacia infinito, de la cantidad

$$s_n = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^n} \qquad [5]$$

En forma más abreviada, pero más expresiva, se puede escribir

$$s_n \to 1 \quad \text{cuando} \quad n \to \infty . \qquad [6]$$

Como otro ejemplo de límite, consideremos las potencias de un número q. Si es $-1 < q < 1$; p. ej., $q = 1/3$ ó $q = -4/5$, la sucesión de potencias de q,

$$q,\ q^2,\ q^3,\ q^4, \ldots,\ q^n, \ldots,$$

se aproxima a cero cuando n crece. Si q es negativo, el signo de q^n será alternativamente $+$ ó $-$, y q^n tenderá a cero aproximándose a

este valor a derecha e izquierda, alternativamente. Para $q = 1/3$ se tiene $q^2 = 1/9$, $q^3 = 1/27$, $q^4 = 1/81$, ..., mientras que si es $q = -1/2$, se tendrá $q^2 = 1/4$, $q^3 = -1/8$, $q^4 = 1/16$, ... Diremos que el *límite de* q^n, *cuando* n *tiende a infinito, es cero*, o, en símbolos

$$q^n \to 0 \quad \text{cuando} \quad n \to \infty, \quad \text{para} \quad -1 < q < 1. \qquad [7]$$

(Incidentalmente, si $q > 1$ ó $q < -1$, q^n no tiende hacia cero, sino que su valor absoluto crece sin límite.)

Para dar una demostración rigurosa de la proposición contenida en [7] partimos de la desigualdad probada en la página 22, la cual decía que $(1 + p)^n \geqslant 1 + np$ para todo entero positivo n y $p > -1$. Si q es un número fijo entre 0 y 1, p. ej., $q = 9/10$, se tiene $q = 1/(1+p)$, siendo $p > 0$. De ahí resulta

$$\frac{1}{q^n} = (1 + p)^n > 1 + np > np,$$

o (véase la regla 4, Cap. VI, suplemento I, 1)

$$0 < q^n < \frac{1}{p} \cdot \frac{1}{n}$$

Por consiguiente, q^n está comprendido entre el valor fijo 0 y $(1/p)\cdot(1/n)$, que se aproxima a cero al crecer n, puesto que p es fijo. De aquí resulta evidentemente que $q^n \to 0$. Si q es negativo, se tiene $q = -1/(1 + p)$ y, en vez de las cotas anteriores 0 y $(1/p)$ $(1/n)$ aparecen ahora, respectivamente, $(-1/p)$ $(1/n)$ y $(1/p)$ $(1/n)$. Por lo demás, el razonamiento es el mismo.

Consideremos ahora la *progresión geométrica*

$$s_n = 1 + q + q^2 + q^3 + \cdots + q^n. \qquad [8]$$

(El caso $q = 1/2$ fué discutido antes.) Como vimos en la página 20, se puede expresar s_n en una forma concisa y simple. Multiplicando s_n por q, encontramos

$$qs_n = q + q^2 + q^3 + q^4 + \cdots + q^{n+1}, \qquad [8a]$$

y restando [8 a] de [8], vemos que desaparecen todos los términos excepto el 1 y el q^{n+1}, obteniendo

$$(1 - q)s_n = 1 - q^{n+1},$$

o, por división,

$$s_n = \frac{1 - q^{n+1}}{1 - q} = \frac{1}{1 - q} - \frac{q^{n+1}}{1 - q}$$

El concepto de límite interviene al hacer crecer n. Como hemos visto, $q^{n+1} = q \cdot q^n$ tiende a cero si es $-1 < q < 1$, y por paso al límite se obtiene

$$s_n \to \frac{1}{1-q} \text{ cuando } n \to \infty, \text{ para } -1 < q < 1. \qquad [9]$$

Escrito como ***progresión geométrica indefinida (serie)*** resulta

$$1 + q + q^2 + q^3 + \ldots = \frac{1}{1-q}, \text{ para } -1 < q < 1. \qquad [10]$$

Por ejemplo,

$$1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \ldots = \frac{1}{1-1/2} = 2,$$

de acuerdo con la ecuación [4], y análogamente

$$\frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3} + \frac{9}{10^4} + \ldots = \frac{9}{10}\,\frac{1}{1-1/10} = 1,$$

de forma que 0,99999 ... = 1. Del mismo modo se tiene que el número decimal finito 0,2374 y el decimal infinito 0,237399999 ... representan el mismo número.

En el capítulo VI daremos un resumen de una exposición general del concepto de límite dentro del moderno espíritu rigorista.

Ejercicios:

1. Pruébese que $1 - q + q^2 - q^3 + q^4 - \ldots = \dfrac{1}{1+q}$, si es $|q| < 1$.

2. ¿Cuál es el límite de la sucesión $a_1, a_2, a_3, \ldots$, donde es $a_n = n/(n+1)$? (*Indicación:* Escríbase el término a_n en la forma $n/(n+1) = 1 - 1/(n+1)$ y obsérvese que el último término del segundo miembro tiende a cero.)

3. ¿Cuál es el límite de $\dfrac{n^2+n+1}{n^2-n+1}$ para $n \to \infty$? (*Indicación:* Póngase la expresión en la forma

$$\frac{1 + \dfrac{1}{n} + \dfrac{1}{n^2}}{1 - \dfrac{1}{n} + \dfrac{1}{n^2}}\;)$$

4. Pruébese, para $|q| < 1$, que $1 + 2q + 3q^2 + 4q^3 + \ldots = \dfrac{1}{(1-q)^2}$.
(*Indicación:* Utilícese el resultado del ejercicio 3 de la página 25.)

5. ¿Cuál es el límite de la serie indefinida

$$1 - 2q + 3q^2 - 4q^3 + \ldots ?$$

6. ¿Cuál es el límite de $\frac{1+2+3+\ldots+n}{n^2}$, de $\frac{1^2+2^2+\ldots+n^2}{n^3}$ y de $\frac{1^3+2^3+\ldots+n^3}{n^4}$? (*Indicación:* Utilícense los resultados de las páginas 19, 21 y 22.)

4. Números racionales y decimales periódicos.—Aquellos números racionales p/q que no son fracciones decimales finitas pueden ser desarrollados en fracciones decimales indefinidas mediante el proceso de división decimal. En cada paso de este proceso debe haber un resto distinto de cero, ya que de otro modo la fracción decimal sería finita. Todos los restos distintos que aparecen en la división son enteros comprendidos entre 1 y $q-1$, de modo que hay solamente $q-1$ posibilidades para los valores de dichos restos. Esto significa que al cabo de q cifras decimales, a lo sumo, algún resto k deberá repetirse. Pero todos los restos siguientes se repetirán en el mismo orden en que aparecían después de este primer resto k. Esto prueba que la *expresión decimal de cualquier número racional es periódica;* una vez aparecido un cierto conjunto finito de cifras, dicho conjunto se repetirá infinitas veces; p. ej., $1/6 = 0{,}166666666\ldots$; $1/7 = 0{,}142857142857142857\ldots$; $1/11 = 0{,}0909090909\ldots$; $122/1100 = 0{,}1109090909\ldots$; $11/90 = 0{,}12222222\ldots$; etc. (Cabe considerar que los números racionales que pueden ser representados mediante fracciones decimales finitas tienen un desarrollo decimal periódico en el cual la cifra 0 se repite indefinidamente después de un número finito de cifras.) En algunos ejemplos anteriores se ve incidentalmente que ciertos desarrollos tienen una parte no periódica, que precede al período que se repite.

Recíprocamente, se puede probar que *todos los decimales periódicos son números racionales.* Como ejemplo, consideremos el decimal periódico indefinido

$$p = 0{,}3322222\ldots.$$

Se tiene $p = 33/100 + 10^{-3}\cdot 2(1 + 10^{-1} + 10^{-2} + \ldots)$, donde la expresión entre paréntesis es la serie geométrica

$$1 + 10^{-1} + 10^{-2} + 10^{-3} + \ldots = \frac{1}{1-1/10} = \frac{10}{9}$$

Por tanto,

$$p = \frac{33}{100} + 2\cdot 10^{-3}\cdot\frac{10}{9} = \frac{2970+20}{9\cdot 10^3} = \frac{2990}{9000} = \frac{299}{900}$$

La demostración en el caso general es esencialmente la misma, pero requiere una notación más general. En el decimal periódico general

$$p = 0{,}a_1a_2a_3\ldots a_mb_1b_2\ldots b_nb_1b_2\ldots b_nb_1b_2\ldots b_n\ldots$$

pongamos $0,b_1b_2 \ldots b_n = B$, de manera que B represente la parte periódica del decimal. Entonces, p puede escribirse así:

$$p = 0, a_1a_2 \ldots a_m + 10^{-m}B(1 + 10^{-n} + 10^{-2n} + 10^{-3n} + \ldots).$$

La expresión entre paréntesis es una serie geométrica con $q = 10^{-n}$; su suma, de acuerdo con la ecuación [10] de la página 74, es $1/(1-10^{-n})$, y, por tanto, se tiene

$$p = 0, a_1a_2 \ldots a_m + \frac{10^{-m}B}{1 - 10^{-n}}$$

Ejercicios:

1. Desarróllense las fracciones $\frac{1}{11}, \frac{1}{13}, \frac{2}{13}, \frac{3}{13}, \frac{1}{17}, \frac{2}{17}$ en fracciones decimales y determínese el período.

*2. El número 142 857 tiene la propiedad de que multiplicado por cualquiera de los números 2, 3, 4, 5 ó 6 da otro que tiene las mismas cifras en diferente orden. Justifíquese esta propiedad, utilizando el desarrollo de 1/7 en fraccion decimal.

3. Desarróllense los números racionales del ejercicio 1 como fracciones «decimales» en los sistemas de numeración de bases 5, 7 y 12.

4. Desarróllese 1/3 como número diádico.

5. Escríbase 0,11212121 . . . como fracción ordinaria. Hállese el valor de dicho símbolo en los sistemas de numeración de base 3 ó 5.

5. Definición general de los números irracionales mediante encajes de intervalos.—En la página 71 adoptamos como definición provisional la siguiente: un *número* es un decimal finito o infinito, y conveníamos en que aquellos decimales infinitos que no correspondían a números racionales serían llamados números irracionales. Sobre la base de los resultados de la sección precedente podemos ahora formular esta definición en la siguiente forma: *el continuo numérico*, o *sistema de números reales* («real» en oposición a los números «imaginarios» o «complejos» que introduciremos en V) es *la totalidad de los decimales infinitos.* (Los decimales finitos quedan considerados como un caso especial de los infinitos: aquel en que todas las cifras a partir de una de ellas son ceros; otra manera de considerarlos será la de reemplazar la última cifra a por $a - 1$ y hacerla seguir de una infinidad de cifras todas iguales a 9. Esto expresa el hecho de que $0,9999 \ldots = 1$, de acuerdo con la sección 3.) Los números *racionales* son los decimales *periódicos*; los números *irracionales* son los decimales *no-periódicos.* Sin embargo, esta definición no es aún plenamente satisfactoria, puesto que, como hemos visto en el capítulo primero, el sistema decimal no se diferencia de los de otras bases por propiedades intrínsecas del sistema de números. En consecuencia, podíamos muy bien haber seguido los razonamientos de las secciones precedentes utilizando el

sistema diádico u otro cualquiera. Por esta razón es de desear una definición más general del continuo numérico, independiente de especiales referencias al sistema de base diez. Quizá la más sencilla sea la siguiente:

Consideremos una sucesión cualquiera de intervalos $I_1, I_2, ..., I_n, ...$ de la recta numérica, cuyos extremos sean puntos racionales, cada uno de los cuales esté contenido en el precedente, y tal que la longitud del intervalo n-ésimo I_n tienda a cero al crecer n. A una tal sucesión la llamaremos *sucesión de intervalos encajados* o *encaje de intervalos*. En el caso de intervalos decimales, la longitud de I_n era 10^{-n}; pero en el caso general puede ser muy bien 2^{-n}, o estar simplemente sujetos a la restricción de ser menores que $1/n$. En estas condiciones vamos a formular un postulado fundamental en geometría: *en correspondencia con cada sucesión de intervalos encajados existe precisamente un punto de la recta numérica que está contenido en todos los intervalos.* (Se ve directamente que no puede haber más de *un* punto común a todos los intervalos, puesto que las longitudes de éstos tienden a cero y dos puntos no pueden estar contenidos en un intervalo de longitud menor que su distancia.) El punto a que se refiere el postulado es por definición un *número real*; si no es un punto racional, se dice que es un *número irracional*. Mediante esta definición establecemos una correspondencia perfecta entre puntos y números. Con ella damos una formulación más general a lo que expresábamos con la definición que utilizaba los decimales infinitos.

Al llegar a este punto quizá el lector se sienta inquietado por una duda completamente legítima: ¿cuál *es* el «punto» de la recta numérica que hemos supuesto pertenecía a todos los intervalos de la sucesión, en el caso de que no se trate de un punto racional? Nuestra respuesta es: la existencia, en la recta numérica, de un punto contenido en cualquier encaje de intervalos cuyos extremos sean puntos racionales es un *postulado fundamental de la geometría*. No es, por tanto, precisa la reducción lógica de esta proposición a otros hechos matemáticos. La aceptamos, lo mismo que aceptamos otros axiomas o postulados de la matemática, a causa de que es plausible de modo intuitivo, y también en virtud de su utilidad para la construcción de un sistema no contradictorio del proceso matemático. Desde un punto de vista puramente formal se podía haber partido de una recta constituída únicamente por puntos racionales y luego *definir* como punto irracional un *símbolo que representa determinadas sucesiones de intervalos racionales encajados*. Un punto irracional queda así perfectamente determinado por la sucesión de intervalos en cuestión. Resulta, por tanto, que nuestro

postulado equivale a una definición. Establecer esta definición después de haber llegado a los encajes de intervalos racionales, gracias al sentimiento intuitivo de que los puntos irracionales «existen», equivale a prescindir de los apoyos intuitivos con los que nuestro razonamiento está acostumbrado a proceder y darse cuenta de que todas las *propiedades matemáticas* de los puntos irracionales pueden venir expresadas como propiedades de los encajes de intervalos racionales.

Tenemos aquí un ejemplo típico de la posición filosófica descrita en la introducción de este libro; prescindir del ingenuo punto de vista «realista», que considera los objetos matemáticos como *cosas en sí* de

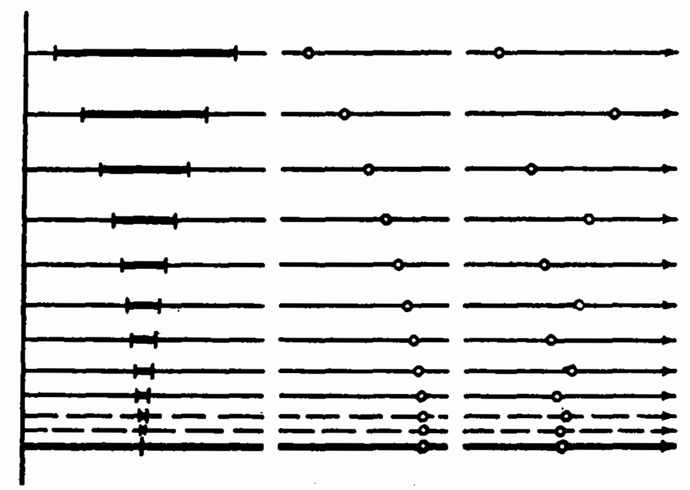


FIG. 11.—Encaje de intervalos. Límites de sucesiones.

las cuales pretendemos modestamente determinar las propiedades, y, en cambio, comprobar que el único modo de existir de los objetos matemáticos que nos importa reside en sus propiedades matemáticas y en las relaciones que los ligan. Estas propiedades y relaciones agotan todos los aspectos posibles bajo los cuales puede intervenir un objeto en el mundo de la actividad matemática. Dejamos de lado la «cosa en sí» matemática, del mismo modo que los físicos dejan de lado el inobservable éter; éste es el significado de la definición *intrínseca* de un número irracional como encaje de intervalos racionales.

La cuestión matemática importante a este respecto es la de que para los números irracionales así definidos, se pueden generalizar de modo inmediato las operaciones de adición, multiplicación, etc., y las relaciones de «menor que» y «mayor que» establecidas para los números racionales, siendo dicha generalización de tal naturaleza que las

reglas y leyes válidas para los números racionales se conservan para los nuevos números; p. ej., la adición de dos números irracionales α y β puede venir dada por medio de los dos encajes de intervalos que sirvieron para definir α y β. Se construye una tercera sucesión de intervalos encajados sumando los valores de los orígenes y los de los extremos de los intervalos correspondientes a los dos encajes que definen α y β, y la nueva sucesión de intervalos encajados define $\alpha + \beta$. Análogamente se pueden definir el producto $\alpha\,\beta$, la diferencia $\alpha - \beta$ y el cociente α/β. Sobre la base de estas definiciones se prueba que las leyes aritméticas que discutimos anteriormente en este capítulo continúan siendo válidas para los números irracionales. Sin embargo, no daremos aquí los detalles de tales demostraciones.

La comprobación de las leyes anteriores es simple y fácil, aunque resulta a veces enojosa para el principiante, que desea más bien aprender las matemáticas como instrumento que analizar sus fundamentos lógicos. Algunos de los textos modernos de matemáticas repelen a numerosos estudiantes, debido a que comienzan con un análisis completo del sistema de los números reales, análisis que resulta un poco pedante. El lector que prescinde simplemente de esa introducción procede con el mismo espíritu con que, hasta fines del siglo pasado, hicieron sus descubrimientos los más grandes matemáticos sobre la base del concepto «ingenuo» del sistema numérico que les proporcionaba su intuición.

Desde el punto de vista físico, la definición de un número irracional mediante un encaje de intervalos corresponde a la determinación de una cantidad observable por una sucesión de medidas de aproximación creciente. Toda operación de determinación, p. ej., de una longitud, tiene significación práctica dentro de los límites de un cierto error posible, error que mide la precisión de la operación. Puesto que los números racionales forman un conjunto denso en la recta, resulta imposible determinar para toda operación física, cualquiera que sea su precisión, si una longitud dada es racional o irracional. De aquí podría parecer que los números irracionales son superfluos para una adecuada descripción de los fenómenos físicos; sin embargo, como veremos más claramente en el capítulo VI, la ventaja efectiva que la introducción de los números irracionales aporta a la descripción matemática de los fenómenos físicos es la de que dicha descripción se simplifica de modo notable, gracias a la posibilidad de utilizar libremente el concepto de límite, para el cual es fundamental el continuo numérico.

***6. Otros métodos de definición de números irracionales. Cortaduras de Dedekind.**—Otra manera de definir los números irracionales

fué dada por Richard Dedekind (1831-1916), uno de los grandes iniciadores del análisis lógico y filosófico de los fundamentos de la matemática. Sus obras *Stetigkeit und irrationale Zahlen* (1872) y *Was sind und was sollen die Zahlen?* (1887) ejercieron una profunda influencia en los estudios sobre los fundamentos de la matemática. Dedekind prefería operar con ideas generales abstractas a hacerlo con sucesiones determinadas de encajes de intervalos. Su método está basado en la definición de «cortadura» que vamos a exponer brevemente.

Supongamos que se ha dado un cierto método para dividir el conjunto de *todos los números racionales* en dos clases, A y B, tales que todo número b de la clase B es mayor que todo elemento a de la clase A. Toda clasificación de este tipo se llama una *cortadura* en el campo de los números racionales. Para una cortadura hay precisamente tres posibilidades, que se excluyen mutuamente:

1) *Hay en* A *un elemento máximo* a*. Este es, p. ej., el caso cuando A está formada por todos los números racionales $\leqslant 1$ y B por todos los números racionales >1.

2) *Hay en* B *un elemento mínimo* b*. Así sucede, p. ej., si A está constituída por todos los números racionales < 1 y B por todos los números racionales $\geqslant 1$.

3) *No hay elemento máximo en* A *ni elemento mínimo en* B. Se tiene, en particular, este caso cuando A está formada por todos los números racionales negativos, el 0, y todos los números racionales positivos cuyo cuadrado es menor que 2, y B por todos los números racionales positivos cuyo cuadrado es mayor que 2. A y B, reunidas, comprenden todos los números racionales, puesto que hemos probado que no hay ningún número racional cuyo cuadrado sea igual a 2.

El caso en que A tuviera un elemento máximo a^* y B un elemento mínimo b^* es imposible, puesto que en dicho caso el número racional $(a^* + b^*)/2$, que está comprendido entre a^* y b^*, sería mayor que el elemento máximo de A y más pequeño que el elemento mínimo de B, y no pertenecería a ninguna de las dos clases.

En el tercer caso, aquel en el que no hay elemento máximo en A ni elemento mínimo en B, la cortadura, según Dedekind, define o simplemente *es* un número irracional. Es fácil ver que esta definición concuerda con la dada mediante encajes de intervalos; cualquier sucesión I_1, I_2, I_3, ... de intervalos encajados define una cortadura, si colocamos en la clase A todos los números racionales que son menores que el origen de al menos uno de los intervalos I_n, y en B todos los demás números racionales.

Desde un punto de vista filosófico, la definición de Dedekind de números irracionales representa un mayor grado de abstracción, puesto que no impone restricción alguna a la ley matemática que define las dos clases A y B. Un método más concreto de definir el continuo numérico real se debe a Georg Cantor (1845-1918). Aunque a primera vista parece completamente diferente del método de los encajes de intervalos y del de las cortaduras, en realidad es equivalente a ambos, en el sentido de que los sistemas numéricos definidos de las tres maneras gozan de las mismas propiedades. La idea de Cantor aparece sugerida por los dos hechos siguientes: 1) los números reales pueden considerarse como decimales de infinitas cifras, y 2) los decimales infinitos son límites de fracciones decimales finitas.

Prescindiendo de la dependencia del sistema decimal se puede decir, siguiendo a Cantor, que toda sucesión $a_1, a_2, a_3, \ldots$ de números racionales define un *número real* si dicha sucesión «converge». La convergencia significa que la diferencia $(a_m - a_n)$ entre dos números cualesquiera de la sucesión tiende a cero cuando a_m y a_n ocupan lugares suficientemente avanzados en la sucesión; es decir, cuando m y n tienden a infinito. (Las sucesivas aproximaciones decimales de todo número gozan de dicha propiedad, puesto que dos números posteriores al n-ésimo difieren en menos de 10^{-n}.) Dado que existen muchas formas de aproximarse al mismo número real mediante sucesiones de números racionales, diremos que dos sucesiones convergentes de números racionales $a_1, a_2, a_3, \ldots$ y $b_1, b_2, b_3, \ldots$ definen el mismo número real si $a_n - b_n$ tiende a cero cuando n crece indefinidamente. Es fácil definir, para tales sucesiones, las operaciones de adición, multiplicación, etc.

III. OBSERVACIONES SOBRE GEOMETRÍA ANALÍTICA [1]

1. El principio fundamental.—El continuo numérico, aceptado como cosa inmediata o bien después de un examen crítico, ha constituído la base de las matemáticas—y en particular de la geometría analítica y del cálculo—desde el siglo XVII.

La introducción del continuo numérico hace posible asociar a cada segmento rectilíneo un determinado número real que da su longitud. Pero aún se puede ir más lejos: no sólo las longitudes, sino también *todo objeto geométrico y toda operación geométrica, pueden ser referidos al reino de los números.* Los pasos decisivos en esta aritmetización de la geometría fueron dados por Fermat (1601-1655) en 1629 y por Descartes (1596-1650) en 1637. La idea fundamental de la geometría analítica es la introducción de «coordenadas»; esto es, de *números* ligados o coordinados con un *objeto geométrico* y que caracterizan completamente a éste. La mayor parte de los lectores conocen, sin duda, las llamadas coordenadas cartesianas rectangulares, que sirven para caracterizar la posición de un punto P en un plano. Se parte de dos rectas fijas perpendiculares del plano, el «eje x» y el «eje y», a las que se refiere todo punto. Estas rectas se consideran como rectas numéri-

[1] El lector que no conozca suficientemente esta materia encontrará en el Apéndice del final del libro una serie de ejercicios sobre los elementos de geometría analítica.

cas orientadas, y se miden con la misma unidad. A cada punto P (Fig. 12) se le asignan dos coordenadas x e y. Estos números se obtienen de la siguiente forma: consideremos el segmento dirigido desde el «origen» O al punto P, y proyectemos este segmento orientado, que a veces se llama «vector de posición» de P, perpendicularmente sobre

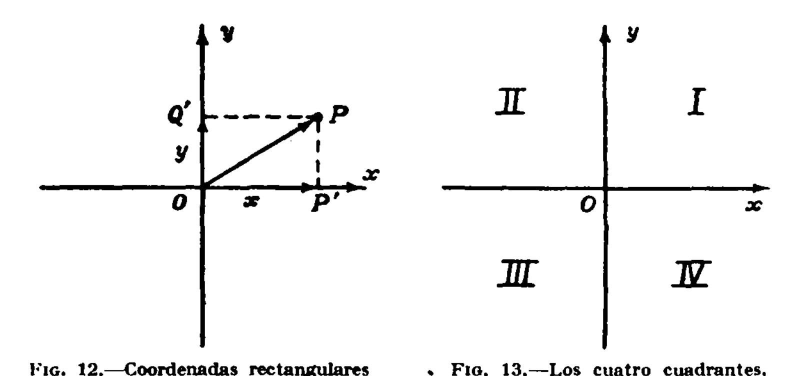


FIG. 12.—Coordenadas rectangulares de un punto.

FIG. 13.—Los cuatro cuadrantes.

los dos ejes; de este modo se obtiene el segmento orientado OP' sobre el eje x, con el número x como medida de su longitud orientada a partir de O, y del mismo modo el segmento orientado OQ' sobre el eje y, con el número y como medida de su longitud orientada a partir de O. Los dos números x, y son las *coordenadas* de P. Recíprocamente: si x, y son dos números arbitrarios dados, el punto P correspondiente está unívocamente determinado. Si x, y son los dos positivos, P está en el *primer cuadrante* del sistema de coordenadas (Fig. 13); si los dos son negativos, P está en el tercer cuadrante; si x es positivo e y negativo, P estará en el cuarto cuadrante y, finalmente, si x es negativo e y positivo, P se hallará en el segundo cuadrante.

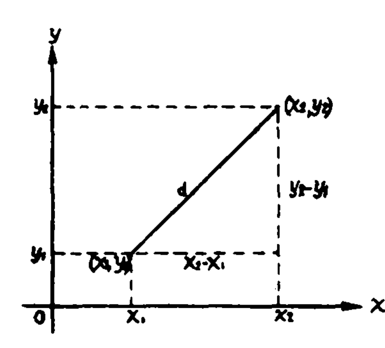


FIG. 14.—Distancia entre dos puntos.

La distancia entre el punto P_1, de coordenadas x_1, y_1, y el punto P_2, de coordenadas x_2, y_2, viene dada por la fórmula

$$d^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2, \qquad [1]$$

la cual se obtiene inmediatamente por el teorema de Pitágoras, como resulta de la figura 14.

***2. Ecuaciones de rectas y curvas.**—Si C es un punto fijo de coordenadas $x = a$, $y = b$, el lugar de todos los puntos P que están a una distancia r dada de C es una circunferencia con centro C y radio r. De la fórmula [1], que da la distancia entre dos puntos, resulta que los puntos de dicha circunferencia tienen coordenadas x, y que satisfacen la ecuación

$$(x - a)^2 + (y - b)^2 = r^2. \qquad [2]$$

Ésta es la llamada *ecuación de la circunferencia*, ya que expresa la condición completa (necesaria y suficiente) que han de cumplir las coordenadas x, y de un punto P para estar sobre la circunferencia de centro C y radio r. Si se desarrollan los paréntesis de [2], dicha ecuación toma la forma

$$x^2 + y^2 - 2ax - 2by = k, \qquad [3]$$

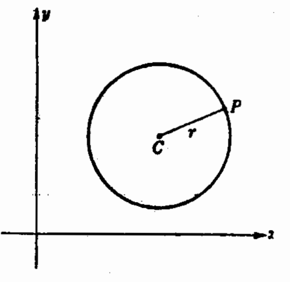


FIG. 15.—La circunferencia.

donde $k = r^2 - a^2 - b^2$. Recíprocamente: si se tiene una ecuación de la forma [3], siendo a, b y k constantes arbitrarias tales que $k + a^2 + b^2$ es positivo, mediante el proceso algebraico de «completar los cuadrados», podemos escribir la ecuación en la forma

$$(x - a)^2 + (y - b)^2 = r^2,$$

donde es $r^2 = k + a^2 + b^2$. Resulta, en consecuencia, que la ecuación [3] define una circunferencia de radio r, que tiene su centro en el punto C de coordenadas a y b.

Las ecuaciones de las rectas tienen una forma todavía más sencilla; p. ej., el eje x tiene como ecuación $y = 0$, puesto que $y = 0$ se verifica para todos los puntos del eje x y sólo para ellos. El eje y tiene la ecuación $x = 0$. Las rectas que pasan por el origen y bisecan los ángulos formados por los ejes tienen como ecuaciones $x = y$ y $x = -y$. Es fácil probar que toda recta tiene una ecuación de la forma

$$ax + by = c, \qquad [4]$$

donde a, b y c son constantes que caracterizan la recta. La significación de la ecuación [4] es la de que todos los pares de valores x, y que la satisfacen son coordenadas de puntos de la recta, y recíprocamente.

Es posible que el lector sepa que la ecuación

$$\frac{x^2}{p^2} + \frac{y^2}{q^2} = 1 \qquad [5]$$

representa una elipse (Fig. 16). Esta curva corta al eje x en los puntos $A(p, 0)$ y $A'(-p, 0)$, y al eje y en $B(0, q)$ y $B'(0, -q)$. (Usaremos la notación $P(x, y)$ o simplemente (x, y) para designar brevemente «el punto P con coordenadas x, y».) Si es $p > q$, el segmento AA', de longitud $2p$, se llama eje mayor de la elipse, mientras que el segmento BB', de longitud $2q$, se llama eje menor. La elipse es el lugar de todos los puntos P cuya suma de distancia a los puntos $F(\sqrt{p^2-q^2}, 0)$

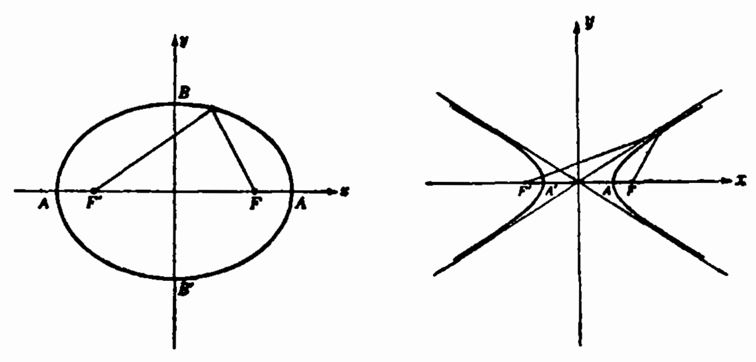


FIG. 16.—La elipse; F y F' son los focos.

FIG. 17.—La hipérbola; F y F' son los focos.

y $F'(-\sqrt{p^2-q^2}, 0)$ es $2p$. El lector puede comprobar esto, como ejercicio, utilizando la fórmula [1]. Los puntos F y F' se llaman *focos* de la elipse, y el cociente $e = \dfrac{\sqrt{p^2-q^2}}{p}$ se denomina *excentricidad* de la misma.

Una ecuación de la forma

$$\frac{x^2}{p^2} - \frac{y^2}{q^2} = 1 \qquad [6]$$

representa una hipérbola. Esta curva (Fig. 17) está formada por dos ramas que cortan al eje x en los puntos $A(p, 0)$ y $A'(-p, 0)$, respectivamente. El segmento AA', de longitud $2p$, se llama eje transversal de la hipérbola. La hipérbola se aproxima indefinidamente a las dos rectas $qx \pm py = 0$ a medida que se aleja del origen, sin llegar nunca a tocarlas; dichas rectas se llaman *asíntotas* de la hipérbola. La

hipérbola es el lugar de todos los puntos P cuya *diferencia* de distancias a los dos puntos $F(\sqrt{p^2 + q^2}, 0)$ y $F'(-\sqrt{p^2 + q^2}, 0)$ es $2p$. Estos dos puntos se llaman focos de la hipérbola; la excentricidad de la curva viene dada por el cociente $e = \dfrac{\sqrt{p^2 + q^2}}{p}$.

La ecuación

$$xy = 1 \qquad [7]$$

define también una hipérbola cuyas asíntotas son los ejes coordenados (Fig. 18). La ecuación de esta hipérbola «equilátera» indica que el área del rectángulo determinado por P y los ejes (Fig. 18) es igual a 1 para todo punto P de la curva.

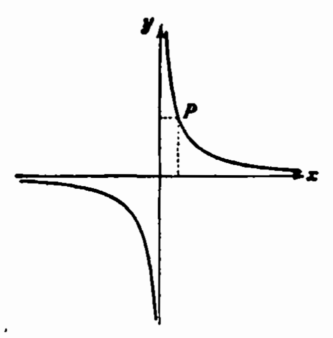


FIG. 18.—La hipérbola equilátera $xy = 1$. El área xy del rectángulo determinado por el punto P (x, y) es igual a 1.

Una hipérbola equilátera cuya ecuación sea

$$xy = c, \qquad [7a]$$

siendo c una constante, es sólo un caso particular de hipérbola, del mismo modo que la circunferencia es un caso particular de la elipse. La particularidad de la hipérbola equilátera reside en el hecho de que sus asíntotas (en el caso presente, los ejes) son perpendiculares entre sí.

Para nosotros, la cuestión fundamental es la idea de que los objetos geométricos pueden ser representados de modo completo mediante elementos aritméticos y algébricos, y que lo mismo sucede con las operaciones geométricas; p. ej., si queremos determinar el punto de intersección de dos rectas, consideraremos las dos ecuaciones

$$\begin{aligned} ax + by &= c \\ a'x + b'y &= c'. \end{aligned} \qquad [8]$$

El punto común a las rectas se hallará determinando sus coordenadas, que no son otra cosa que la solución x, y del sistema [8]. Análogamente, los puntos de intersección de dos líneas, tales como la circunferencia $x^2 + y^2 - 2ax - 2by = k$ y la recta $ax + by = c$, se hallarían resolviendo el sistema formado por las ecuaciones de ambas líneas.

IV. ANÁLISIS DEL CONCEPTO MATEMÁTICO DE INFINITUD

1. Conceptos fundamentales.—La sucesión de los enteros positivos

$$1, 2, 3, \ldots$$

es el primer y mas importante ejemplo de conjunto infinito. No hay ningún misterio en el hecho de que esta sucesión no tiene fin; puesto que, por grande que sea el entero n, siempre se puede formar el entero siguiente $n + 1$. Pero el paso del *adjetivo* «infinito», que significa simplemente «sin fin», al *sustantivo* «infinito» no debe hacernos pensar que «infinito», representado generalmente con el símbolo ∞, puede considerarse como si fuera un *número* ordinario. No es posible incluir el símbolo ∞ en el sistema de los números reales y conservar al mismo tiempo las leyes fundamentales de la aritmética. Sin embargo, el concepto de infinito invade toda la matemática, ya que los objetos matemáticos son considerados habitualmente, no como individuos, sino como miembros de clases o conjuntos que contienen una infinidad de objetos del mismo tipo, tales como el conjunto de todos los enteros, o el de los números reales, o el de los triángulos de un plano. Por esta razón es necesario analizar el infinito matemático de una manera precisa. La teoría moderna de conjuntos, creada por Georg Cantor y su escuela a fines del siglo pasado, obtuvo brillantes resultados en tal cometido. La teoría de conjuntos de Cantor ha penetrado y ejercido enorme influjo en varios campos de la matemática, y ha llegado a ser de importancia fundamental en el estudio de las bases lógicas y filosóficas de dicha ciencia. El punto de partida es el concepto general de *conjunto* o *agregado*. En este concepto se comprende toda colección de objetos definida por una regla determinada que especifica exactamente qué objetos pertenecen a la colección. Como ejemplos podemos considerar el conjunto de todos los enteros positivos; el de todos los decimales periódicos; el conjunto de todos los números reales, o el de todas las rectas del espacio de tres dimensiones.

Para comparar la «magnitud» de dos conjuntos diferentes es fundamental la noción de «equivalencia». Si los elementos de dos conjuntos A y B pueden ser apareados, de modo que a todo elemento de A corresponda un elemento (y sólo uno) de B y a cada elemento de B corresponda un elemento de A, y uno solo, la correspondencia entre A y B se llama *biunívoca*, y dichos conjuntos se dicen *equivalentes* o *coordinables*. La noción de equivalencia para conjuntos *finitos* coincide con la noción ordinaria de *igualdad de números*, puesto que dos con-

juntos finitos tienen el mismo número de elementos cuando (y sólo entonces) se pueden poner en correspondencia biunívoca. Ésta es, en realidad, la verdadera idea que interviene en la operación de contar, ya que cuando contamos un conjunto finito de objetos lo que hacemos es establecer una correspondencia biunívoca entre dichos objetos y el conjunto de símbolos numéricos 1, 2, 3, ... , n.

No es necesario siempre contar los objetos de dos conjuntos finitos para comprobar su equivalencia; p. ej., se puede afirmar, sin necesidad de contar, que todo conjunto finito de circunferencia de radio 1 es coordinable con el conjunto de sus centros.

La idea de Cantor fué la de extender el concepto de equivalencia a los conjuntos infinitos para poder de este modo definir una «aritmética» de los infinitos. El conjunto de todos los números reales y el conjunto de todos los puntos de una recta son equivalentes, puesto que la elección del origen y del segmento unidad permite asociar de modo biunívoco cada punto P de la recta con un número real x determinado, su abscisa:

$$P \longleftrightarrow x.$$

Los *enteros pares* forman un subconjunto propio del conjunto de *todos los enteros*, y los *enteros* forman un subconjunto propio del conjunto de todos los *números racionales*. (Con la frase *subconjunto propio* de un conjunto S designamos un conjunto S' formado por algunos, *no todos*, de los objetos de S). Evidentemente, *si un conjunto es finito*, es decir, si contiene un cierto número n de elementos, *no puede ser equivalente a ninguno de sus subconjuntos propios*, ya que cualquier subconjunto puede contener a lo sumo $n-1$ elementos. En cambio, *si un conjunto contiene infinitos objetos* se verifica, lo que a primera vista parece paradójico, *que es equivalente a algunos de sus subconjuntos propios;* p. ej., la coordinación

$$\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 \ldots & n \ldots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\ 2 & 4 & 6 & 8 & 10 \ldots & 2n \ldots \end{array}$$

establece una correspondencia biunívoca entre el conjunto de los *enteros positivos* y su subconjunto propio formado por los *enteros pares*, y, por tanto, ambos conjuntos son equivalentes. Esta contradicción con el familiar axioma «el todo es mayor que cualquiera de sus partes» muestra qué sorpresas pueden presentarse en el dominio del infinito.

2. La numerabilidad de los números racionales y la no-numerabilidad del continuo.—Uno de los primeros descubrimientos de Cantor

en su análisis del infinito fué el de que el conjunto de los *números racionales* (que contiene el conjunto infinito de los enteros y que, por consiguiente, es infinito) es equivalente al *conjunto de los enteros*. A primera vista resulta bastante extraño que el conjunto denso en la recta de los números racionales pueda ser equivalente al disperso conjunto de los enteros. Ciertamente, no se pueden colocar los números positivos racionales *en orden de magnitud* (como ocurre con los enteros), de modo que haya un a que sea el primer número racional, b el inmediato mayor, y así sucesivamente, puesto que existen infinitos números racionales entre dos cualesquiera dados, y, por consiguiente, carece de sentido la expresión «el inmediato mayor». Pero, como observó Cantor, prescindiendo de la relación de magnitud entre los elementos consecutivos, es posible colocar los números racionales en una sucesión r_1, r_2, r_3, ..., análoga a la de los enteros. En dicha sucesión habrá un primer número racional, un segundo, un tercero y así sucesivamente, y todo número racional aparecerá precisamente una vez. Una ordenación de los elementos de un conjunto en una sucesión análoga a la de los enteros se llama una *enumeración* del conjunto. Construyendo una enumeración del conjunto de los números racionales, Cantor probó que dicho conjunto es equivalente al de los números enteros, ya que la correspondencia

$$\begin{array}{ccccccc} 1 & 2 & 3 & 4 & \cdots & n & \cdots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \updownarrow & \\ r_1 & r_2 & r_3 & r_4 & \cdots & r_n & \cdots \end{array}$$

es biunívoca. Vamos a describir una manera efectiva de enumeración de los números racionales.

Todo número racional puede escribirse en la forma a/b, donde a y b son enteros, y todos aquellos números pueden disponerse en un cuadro, con a/b en la fila a y en la columna b; p. ej., 3/4 estará en la tercera fila y en la cuarta columna. Entonces, todos los números racionales positivos pueden ser ordenados según el esquema siguiente: en el cuadro anterior trazamos una línea quebrada continua que pase por todos los números del cuadro en la forma indicada en la figura 19. Partiendo del 1, vamos horizontalmente hasta el primer lugar a la derecha, obteniendo el 2 como segundo término de la sucesión; vamos entonces diagonalmente hacia abajo, a la izquierda, hasta la primera columna, a alcanzar la posición ocupada por el 1/2; luego, verticalmente, hacia abajo, un lugar, hasta el 1/3; después, diagonalmente hacia arriba y a la derecha, alcanzamos el 3: de ahí pasamos al 4 y después, diagonalmente, hasta el 1/4, y así sucesivamente. A través de esta línea que-

brada se obtiene la sucesión 1, 2, 1/2, 1/3, 2/2, 3, 4, 3/2, 2/3, 1/4, 1/5, 2/4, 3/3, 4/2, 5. ... que contiene los números racionales en el orden en que aparecen en la línea quebrada. En dicha sucesión suprimimos aquellos números a/b para los cuales a y b tienen un factor común, de modo que cada número racional aparezca una sola vez, representado por una fracción irreducible. Se obtiene así la sucesión

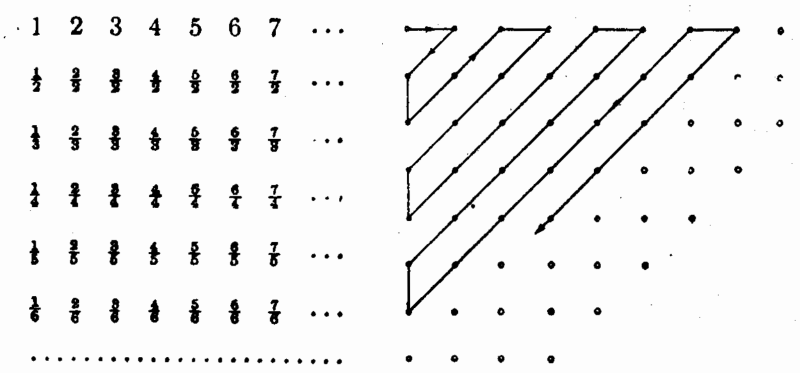


FIG. 19.—El conjunto de los números racionales es numerable.

1, 2, 1/2, 1/3, 3, 4, 3/2, 2/3, 1/4, 1/5, 5, ... la cual contiene cada número racional una sola vez. Resulta así la numerabilidad del conjunto de los números racionales. Como consecuencia de la correspondencia biunívoca entre los números racionales y los puntos racionales de una recta, se tiene la numerabilidad del conjunto de los puntos racionales de la recta.

Ejercicios:

1. Pruébese que el conjunto de todos los enteros positivos y negativos es numerable. Pruébese que el conjunto de los números racionales positivos y negativos es numerable.

2. Pruébese que el conjunto $S + T$ (véase pág. 120) es numerable si S y T son numerables. Pruébese el mismo hecho para el caso de una suma de tres, cuatro o cualquier número n de conjuntos, y, finalmente, para un conjunto numerable de conjuntos numerables.

Habiendo demostrado que el conjunto de números racionales es numerable, podría esperarse que *todo* conjunto infinito fuera también numerable, con lo que se tendría el resultado final del análisis del infinito. La realidad de los hechos es muy distinta. El mismo Cantor obtuvo el notable resultado de que el *conjunto de los números reales*, racionales e irracionales, *no es numerable*. En otros términos, el conjunto de los números reales presenta una diferencia radical y, por así decirlo, un tipo superior de infinitud que el de los enteros o el de los

números racionales. La ingeniosa demostración indirecta de este hecho, debida a Cantor, ha quedado como modelo para un cierto tipo de demostraciones matemáticas. Las líneas principales de dicha demostración son las siguientes: se parte de la hipótesis de que todos los números reales pueden ser efectivamente enumerados y dispuestos en una sucesión, y se construye luego un número real que no figura en dicha sucesión. De ahí resulta una contradicción con la hipótesis inicial, que suponía que *todos* los números reales estaban incluídos en la sucesión; la hipótesis de que la enumeración de los números reales es posible resulta absurda y, en consecuencia, la hipótesis contraria, es decir, la proposición de Cantor de que el conjunto de los números reales no es numerable queda demostrada.

Para llevar a cabo el programa indicado, supongamos que hemos enumerado todos los números reales colocándolos en una tabla de decimales infinitos

1.er número	$N_1,a_1a_2a_3a_4a_5 \cdots$
2.º número	$N_2,b_1b_2b_3b_4b_5 \cdots$
3.er número	$N_3,c_1c_2c_3c_4c_5 \cdots$
. . . . . .	. . . .

donde las N designan la parte entera y las letras minúsculas las cifras decimales. Supongamos que esta sucesión de fracciones decimales contiene *todos* los números reales. El punto esencial de la demostración consiste en construir, por un «proceso diagonal», un número del que se demuestra que no puede estar en la sucesión. Para ello, comencemos por elegir una cifra a distinta de a_1 y de 0 y 9 (las últimas exigencias se hacen para evitar ambigüedades que puedan resultar de igualdades análogas a la $0{,}9999 \ldots = 1{,}0000 \ldots$); luego, una cifra b distinta de b_2, de 0 y de 9; después, una c diferente de c_3, de 0 y de 9, y así sucesivamente. (P. ej., se puede elegir $a = 1$, a no ser que sea $a_1 = 1$, en cuyo caso tomaríamos $a = 2$, y análogamente para las cifras $b, c, d, e, \ldots$.) Consideremos entonces el decimal infinito

$$z = 0{,}abcde \cdots .$$

Este nuevo número z es ciertamente distinto de cualquiera de los que están en la tabla anterior; no puede ser igual al primero, puesto que difiere de él en la primera cifra decimal; no puede ser igual al segundo, porque tiene distinta la segunda cifra decimal; y, en general, no puede ser igual al que ocupa el lugar n, porque difiere de él en la n-ésima cifra decimal. Esto prueba que dicha tabla, en la que hemos dispuesto una sucesión de números reales, *no* puede contener todos los números reales. El conjunto de estos números no es, por tanto, numerable.

Quizá pueda sospechar el lector que la razón de la no-numerabilidad del continuo numérico esté en el hecho de que la recta se extiende infinitamente, y que posiblemente pueda ser numerable el conjunto de los puntos de un segmento finito de recta. Se comprueba fácilmente que no es éste el caso viendo que el continuo numérico es equivalente a cualquier segmento finito; p. ej., al (0,1), del que se han excluído los extremos. La correspondencia biunívoca deseada se puede obtener formando con dicho segmento una quebrada de tres lados de

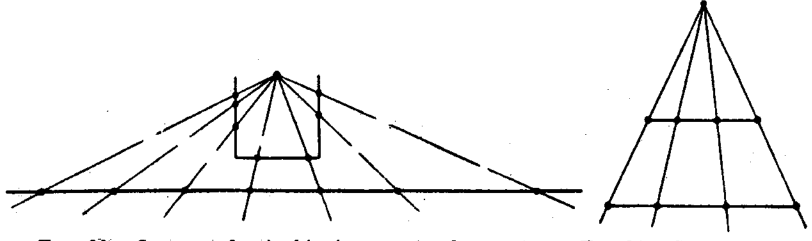


FIG. 20.—Correspondencia biunívoca entre los puntos de una poligonal y los de la recta completa.

FIG. 21.—Correspondencia biunívoca entre los puntos de dos segmentos de distinta longitud.

longitud igual a 1/3 y proyectando desde un punto conveniente, como se indica en la figura 20. Resulta, por tanto, que también un segmento finito de la recta numérica contiene una infinidad no-numerable de puntos.

Ejercicio: Pruébese que cualquier intervalo $[A, B]$ de la recta numérica es equivalente a otro intervalo cualquiera $[C, D]$.

Vale la pena indicar otra demostración, quizá más intuitiva, de la no-numerabilidad del continuo numérico. Apoyándonos en lo que acabamos de decir, será suficiente que nos ocupemos del conjunto de puntos comprendidos entre 0 y 1. La nueva demostración será también por reducción al absurdo. Supongamos que el conjunto de todos los puntos de la recta comprendidos entre 0 y 1 puede ser ordenado en una sucesión

$$a_1, a_2, a_3, \cdots . \qquad [1]$$

Incluyamos el punto a_1 en un intervalo de longitud 1/10, el punto a_2 en un intervalo de longitud $1/10^2$, y así sucesivamente. Si todos los puntos del intervalo (0,1) estuvieran en la sucesión [1], el intervalo unidad aparecería completamente recubierto, en parte quizá por intervalos superpuestos, por la sucesión de intervalos de longitudes 1/10, $1/10^2$, ... que hemos construído. (El hecho de que alguno de los inter-

valos pueda extenderse fuera del segmento unidad no influye en la demostración. La suma de las longitudes de dichos intervalos viene dada por la serie geométrica

$$1/10 + 1/10^2 + 1/10^3 + \dots = \frac{1}{10}\left[\frac{1}{1-\frac{1}{10}}\right] = \frac{1}{9}$$

Así, la hipótesis de que la sucesión [1] contiene todos los números reales comprendidos entre 0 y 1 que llenan el intervalo completo de longitud 1 conduce a la contradicción de que dicho intervalo puede ser recubierto por un conjunto de intervalos de longitud total igual a 1/9, lo cual es intuitivamente absurdo. Aceptaremos esta contradicción como una demostración, aunque desde un punto de vista lógico requeriría un análisis más detallado.

El razonamiento del párrafo anterior sirve para establecer un teorema de gran importancia en la teoría moderna de la *medida*. Reemplazando los intervalos anteriores por otros más pequeños de longitud $\varepsilon/10^n$, donde ε es un número positivo arbitrariamente pequeño, se ve que los puntos de todo conjunto numerable de la recta pueden ser incluídos en un conjunto de intervalos de longitud total igual a $\varepsilon/9$. Puesto que ε es arbitrario, dicha longitud total puede ser tan pequeña como se quiera. En la terminología de la teoría de la medida expresaremos este hecho diciendo que un conjunto numerable de puntos tiene *medida nula*.

Ejercicio: Demuéstrese que el mismo resultado tiene lugar para un conjunto numerable de puntos de un plano, reemplazando las longitudes de los intervalos por áreas de cuadrados.

3. «Números cardinales» de Cantor.—Resumiendo los resultados expuestos hasta ahora, se tiene: el número de elementos de un conjunto *finito* A no puede ser igual al número de elementos de un conjunto finito B, si A contiene *más* elementos que B. Si reemplazamos el concepto de «conjuntos con el mismo número (finito) de elementos» por el concepto más general de *conjuntos equivalentes*, se tiene que para conjuntos infinitos no es válida la proposición anterior; el conjunto de todos los enteros contiene más elementos que el conjunto de los números pares, y el conjunto de los números racionales más que el de enteros y, sin embargo, estos tres conjuntos son equivalentes. Esto podría hacer sospechar que *todos* los conjuntos infinitos fueran equivalentes y que toda otra distinción que la de conjuntos finitos e infinitos resultaría superflua; pero hemos visto que los resultados de Cantor contradicen tal sospecha; existe un conjunto, el continuo numérico real, que no es equivalente a ningún conjunto numerable.

Así resulta que hay al menos dos tipos diferentes de «infinitud»: el infinito numerable de los enteros y el infinito no-numerable del con-

tinuo. Si dos conjuntos, finitos o infinitos, son equivalentes, diremos que tienen el *mismo número cardinal.* Esta propiedad se reduce a la noción ordinaria de *igual número natural* si A y B son finitos, y puede considerarse como una generalización de este concepto. Por otra parte, si un conjunto A es equivalente con algún subconjunto de B, mientras que B no es equivalente con A ni con ninguno de sus subconjuntos, diremos, siguiendo a Cantor, que el conjunto B tiene un *número cardinal mayor* que el del conjunto A. El uso de la palabra «número» está también en este caso de acuerdo con la noción de «número mayor» para conjuntos finitos. El conjunto de los enteros es un subconjunto del conjunto de los números reales, mientras que el conjunto de los números reales no es equivalente con el de los enteros ni con ninguno de los subconjuntos de éste (es decir, el conjunto de los números reales no es numerable ni finito); por tanto, de acuerdo con nuestra definición, el continuo de los números reales tiene un número cardinal mayor que el conjunto de los enteros.

*Se debe indicar que Cantor demostró realmente cómo construir una sucesión de conjuntos infinitos de números cardinales crecientes. Partiendo de los números enteros positivos, es suficiente probar que *dado un conjunto* A *cualquiera, es posible construir otro conjunto* B *de mayor número cardinal que aquél.* A causa de la gran generalidad de este teorema su demostración resulta bastante abstracta. Definimos el conjunto B como el conjunto de todos los distintos subconjuntos de A. En el término «subconjunto» debemos incluir no solamente los subconjuntos propios de A, sino también A mismo y el «subconjunto vacío», 0, que no contiene ningún elemento. (Así, en el caso en que A esté formado por los enteros 1, 2, 3, B consta de los 8 elementos diferentes $\{1, 2, 3\}$, $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$, $\{1\}$, $\{2\}$, $\{3\}$ y 0.) Los elementos de B son a su vez *conjuntos* formados por elementos de A. Supongamos que B es equivalente a A o a algún subconjunto de A; es decir, que existe una ley que coordina de manera biunívoca los elementos de A o de uno de sus subconjuntos con todos los elementos de B; esto es, con los subconjuntos de A:

$$a \longleftrightarrow S_a, \qquad [2]$$

donde designamos con S_a el subconjunto de A que corresponde al elemento a de A. Llegaremos a una contradicción construyendo un elemento de B (es decir, un subconjunto de A) al que no le corresponde ningún elemento a de A. En la construcción de este subconjunto se observa que existen dos posibilidades para todo elemento x de A: o bien el conjunto S_x asignado a x por la correspondencia [2] contiene al elemento x, o S_x no contiene a x. *Definamos* T *como el subconjunto de* A *formado por los elementos* x *tales que* S_x *no contiene a* x. Este subconjunto T difiere de cualquier S_a al menos en el elemento a, puesto que si S_a contiene a a, T no debe contenerlo, mientras que si S_a no contiene a a, T debe contenerlo. En consecuencia, T no está incluído en la correspondencia [2]. Esto prueba que no se puede establecer una correspondencia biunívoca entre los elementos de A, o de un subconjunto de A, y los de B. Sin embargo, la coordinación

$$a \longleftrightarrow \{a\}$$

define una correspondencia biunívoca entre los elementos de A y el subconjunto de B formado por los subconjuntos de A que constan de un solo elemento. En consecuencia, de acuerdo con la definición del último párrafo, B tiene un número cardinal mayor que el de A.

*Ejercicio: Si A contiene n elementos, siendo n un entero positivo, pruébese que B, tal como ha sido definido antes, contiene 2^n elementos. Si A es el conjunto de todos los enteros positivos, demuéstrese que B es equivalente al continuo numérico entre 0 y 1. (*Indicación:* Represéntese todo subconjunto de A en el primer caso por una sucesión finita y en el segundo por una sucesión infinita de las cifras 0 y 1,

$$a_1 a_2 a_3 \cdots$$

siendo $a_n = 1$ ó 0, según que el n-ésimo elemento de A pertenezca o no al subconjunto en cuestión.)

Podría pensarse que es fácil encontrar un conjunto de *puntos* con número cardinal mayor que el conjunto de los números reales entre 0 y 1. Puesto que, a primera vista, un cuadrado, siendo de «dimensión 2», parece contener «más» puntos que un segmento «unidimensional». Paradójicamente, las cosas no ocurren de ese modo: *el número cardinal del conjunto de puntos de un cuadrado es el mismo que el del conjunto de puntos de un segmento.* Para probar esta afirmación, estableceremos la correspondencia siguiente:

Si (x, y) es un punto de un cuadrado de lado unidad, x e y pueden ser escritos en forma decimal

$$x = 0{,}a_1a_2a_3a_4 \cdots ,$$
$$y = 0{,}b_1b_2b_3b_4 \cdots ,$$

donde para evitar ambigüedades escribiremos, p. ej., 0,250000 . . . en vez de 0,249999 · · · para el número racional 1/4. Al punto (x, y) del cuadrado le hacemos corresponder el punto

$$z = 0{,}a_1b_1a_2b_2a_3b_3a_4b_4 \cdots$$

del segmento entre 0 y 1. Evidentemente, a puntos distintos (x, y) y (x', y') del cuadrado les corresponderán puntos distintos z y z' del segmento, de modo que el número cardinal del cuadrado no puede exceder del cardinal del segmento.

(En realidad, la correspondencia que acabamos de establecer es biunívoca entre el conjunto de todos los puntos del cuadrado y un subconjunto propio del segmento unidad; puesto que ningún punto del cuadrado corresponde al punto 0,21409090909 · · · del segmento, ya que, como hemos indicado, tomamos la forma 0,250000 · · · y no la 0,2499999 · · · para representar el número 1/4. Sin embargo, es posible modificar un poco la correspondencia, de modo que se tenga la biunivocidad entre el cuadrado y el segmento, los cuales resultan así con el mismo número cardinal.)

Un razonamiento análogo al anterior mostraría que el número cardinal de los puntos de un cubo es igual al número cardinal del segmento unidad.

Aunque estos resultados parecen contradecir la noción intuitiva de dimensión, debemos recordar que la correspondencia que hemos definido no es «continua»; si recorremos el segmento de 0 a 1 de modo continuo, los puntos correspondientes del cuadrado no forman una curva continua, sino que aparecen en un orden completamente caótico. La dimensión de un conjunto de puntos depende, no solamente del número cardinal del conjunto, sino también del modo como los puntos

aparecen distribuídos en el espacio. En el capítulo V volveremos a ocuparnos de esta cuestión.

4. El método de demostración indirecta (demostraciones por reducción al absurdo).—La teoría de los números cardinales no es más que un aspecto de la teoría general de conjuntos, creada por Cantor en lucha con las severas críticas de algunos de los más distinguidos matemáticos de su tiempo. Algunos de estos críticos, tales como Kronecker y Poincaré, le reprochaban la vaguedad del concepto general de «conjunto» y el carácter no constructivo de los razonamientos utilizados para definir algunos conjuntos.

Las objeciones a los razonamientos no constructivos se refieren a las que podemos llamar *pruebas indirectas* o *demostraciones por reducción al absurdo*. Las demostraciones indirectas constituyen un tipo habitual de razonamiento matemático: para establecer la verdad de una proposición A, se hace la hipótesis de que la proposición A', contraria de la A, es cierta. Entonces, mediante una cadena de razonamientos, se llega a una contradicción con A', lo que prueba lo absurdo de A'. En consecuencia, apoyándose en el principio lógico fundamental del *tertio excluso*, la falsedad de A' establece la verdad de A.

A lo largo de este libro encontraremos ejemplos de demostraciones indirectas que pueden convertirse en demostraciones directas, si bien la forma indirecta presenta las ventajas de la brevedad y de prescindir de detalles que no son necesarios para el objetivo inmediato. Existen, sin embargo, teoremas para los cuales no ha sido posible dar más demostración que la indirecta; aún más: hay teoremas que se pueden demostrar por el método indirecto, para los cuales no sería posible ni siquiera en principio dar una demostración directa constructiva, a causa de la naturaleza misma del teorema. Tal es, p. ej., el teorema dado en la página 90. En distintas ocasiones de la historia de las matemáticas, en las que los esfuerzos de los matemáticos se dirigían hacia la obtención de soluciones *constructivas* para determinados problemas, con vistas a mostrar la posibilidad de solución, se llegó a la construcción gracias a haberse encontrado una demostración indirecta y no constructiva.

Existe una diferencia esencial entre probar la existencia de un objeto de cierto tipo mediante la construcción de un ejemplo tangible de tal objeto y demostrar que la no existencia de dicho objeto conduciría a una contradicción. En el primer caso se tiene un objeto tangible, mientras que en el segundo no se tiene más que una contradicción. Algunos matemáticos destacados han propugnado en tiempos recientes la supresión más o menos completa de las demostraciones

no constructivas de la matemática. Aun en el caso en que se considere deseable este programa, en el estado actual de la matemática supondría una gran complicación y también la parcial destrucción del edificio matemático actual. Por esta razón no es de extrañar que la escuela «intuicionista», que es la que ha adoptado tal programa, haya encontrado fuerte resistencia, y que los intuicionistas más puros no puedan en ocasiones permanecer fieles a sus principios.

5. Las paradojas del infinito.—Aunque la posición irreducible de los intuicionistas parece demasiado extremista a la mayor parte de los matemáticos, la aparición de paradojas lógicas en la teoría de los conjuntos infinitos representó una seria amenaza para dicha teoría. Pronto se observó que la utilización sin restricciones del concepto de «conjunto» podía conducir a contradicciones. Una de las paradojas, presentada por Bertrand Russell, puede ser formulada como sigue: la mayor parte de los conjuntos no se contienen a sí mismos como elementos; p. ej., el conjunto A de los números enteros contiene como elementos únicamente números enteros; como A no es un entero, sino un *conjunto de enteros,* no se contiene a sí mismo como elemento. Un conjunto de este tipo se llamará «ordinario». Sin embargo, es posible que un conjunto se contenga a sí mismo como elemento; p. ej., el conjunto S, definido por la frase «S contiene como elementos los conjuntos que se pueden definir en castellano con una frase de menos de treinta palabras», es un conjunto que se contiene a sí mismo como elemento. A estos conjuntos los llamaremos «extraordinarios». En todo caso, la mayor parte de los conjuntos son ordinarios y podemos excluir el comportamiento extraño de los conjuntos «extraordinarios», limitando nuestra atención al *conjunto de todos los conjuntos ordinarios.* Llamemos C a dicho conjunto. Cualquier elemento de C es a su vez un conjunto; en realidad un conjunto ordinario. Ahora se plantea la cuestión de saber si el conjunto C es ordinario o extraordinario. Debe ser de uno o del otro tipo; si C es ordinario, debe contenerse a sí mismo como elemento, puesto que hemos definido C por la propiedad de contener a *todos* los conjuntos ordinarios; pero si es así, C debe ser extraordinario, ya que hemos llamado extraordinarios a los conjuntos que se contienen a sí mismos como elementos. Se tiene en consecuencia una contradicción; por tanto, C debe ser extraordinario. Pero entonces C, que se contiene a sí mismo por ser extraordinario, contendrá un conjunto extraordinario (el mismo C), en contradicción con la definición que dimos de C como conjunto que contiene únicamente conjuntos ordinarios. Vemos así que, en cualquiera de las dos hipótesis, la mera existencia de C conduce a una contradicción.

6. Los fundamentos de la matemática.—Paradojas análogas a la precedente condujeron a Russell y a otros matemáticos a un estudio sistemático de los fundamentos de la matemática y de la lógica. El objeto final de sus esfuerzos es el de dar al razonamiento matemático una base lógica, la cual se pueda probar que está libre de contradicción, y que al mismo tiempo incluya todo lo que es considerado como importante por todos (o algunos de) los matemáticos. Aunque esta meta ambiciosa no ha sido alcanzada y quizá no pueda ser alcanzada nunca, el tema de la lógica matemática ha atraído la atención de un número creciente de estudiosos. Muchas de las cuestiones en este dominio que pueden ser enunciadas en forma simple presentan grandes dificultades para su solución. Como ejemplo, citaremos la llamada *hipótesis del continuo*, que supone la no existencia de ningún conjunto cuyo número cardinal sea mayor que el de los conjuntos numerables y menor que el del continuo numérico correspondiente al conjunto de los números reales. De esta hipótesis pueden deducirse muchas consecuencias interesantes; pero hasta ahora no ha sido demostrada ni refutada, si bien recientemente Kurt Gödel ha probado que si el sistema de postulados sobre los que se funda la teoría de conjuntos no es contradictorio, tampoco lo es el sistema ampliado obtenido al añadir la hipótesis del continuo. Cuestiones tales como éstas se reducen en última instancia a saber lo que se quiere significar por el concepto de *existencia matemática*. Afortunadamente, la existencia de la matemática no depende de una respuesta satisfactoria. La escuela de los «formalistas», dirigida por el gran matemático Hilbert, afirma que, en matemática, «existencia» significa simplemente «libre de contradicción». Resulta entonces necesario construir un sistema de postulados del que pueda deducirse toda la matemática por razonamiento puramente formal, y demostrar que este sistema de postulados no puede llevar nunca a contradicción. Los resultados recientes obtenidos por Gödel y otros parecen probar que este programa, al menos como fué concebido originalmente por Hilbert, no puede realizarse. Es significativo que la teoría de Hilbert acerca de la estructura formal de las matemáticas esté basada en esencia en un proceso intuitivo. De una forma u otra, explícita o implícitamente, aun bajo el más inflexible aspecto formalista, lógico o postulatorio, la intuición constructiva continúa siendo el elemento vital en matemáticas.

V. NÚMEROS COMPLEJOS

1. Origen de los números complejos.—Por muchas razones el concepto de número ha tenido que ser extendido más allá del continuo

numérico real mediante la introducción de los llamados *números complejos*. Debe observarse que en el desarrollo histórico y psicológico de las matemáticas, todas estas generalizaciones y nuevas invenciones no han sido en forma alguna resultado de algún esfuerzo individual, sino que más bien aparecen como el desenlace de una evolución gradual y cautelosa que no puede atribuirse a una sola persona determinada. Fué la necesidad de una mayor libertad en el cálculo formal lo que llevó a la utilización de los números racionales negativos, y sólo al final de la Edad Media empezaron los matemáticos a perder el temor de usar estos conceptos que no tenían el mismo carácter concreto e intuitivo de los números naturales. Hasta la mitad del siglo XIX los matemáticos no percibieron de una forma completamente clara que la base esencial lógico-filosófica de las operaciones en un conjunto numérico ampliado es formalista, y que las ampliaciones han de hacerse mediante definiciones que, como tales, son libres, pero resultan inútiles si no son hechas de manera que las leyes y propiedades válidas en el campo numérico original se conserven al ampliar éste. El hecho de que estas ampliaciones puedan estar a veces relacionadas con objetos «reales», y que de esta forma procuren una herramienta para nuevas aplicaciones, es de la mayor importancia, si bien esto es sólo una justificación, pero no constituye una prueba lógica de la validez de la ampliación.

El primer problema que requiere el uso de los números complejos es el de la *resolución de las ecuaciones cuadráticas*. Recordemos el concepto de ecuación lineal $ax = b$, en la cual hay que determinar la incógnita x. La solución es simplemente $x = \frac{b}{a}$, y la exigencia de que toda ecuación de coeficientes enteros $a \neq 0$ y b tenga solución, precisa de la introducción de los números racionales. Ecuaciones tales como

$$x^2 = 2, \qquad [1]$$

que carecen de solución en el campo de los números racionales, nos llevan a construir el campo más amplio de los números reales, en el cual existe solución. Pero incluso el campo de los números reales no es suficientemente amplio para procurarnos una teoría completa de las ecuaciones cuadráticas. Una ecuación sencilla tal como

$$x^2 = -1 \qquad [2]$$

carece de solución real, pues el cuadrado de cualquier número real no es nunca negativo.

Podemos contentarnos con la afirmación de que esta sencilla ecua-

ción no es resoluble, o bien seguir el camino usual de extender nuestro concepto de número mediante la introducción de números que permitan resolver la ecuación. Exactamente esto es lo que se hace al introducir el nuevo símbolo i por la definición $i^2 = -1$. Naturalmente, este objeto i, la «unidad imaginaria», no tiene nada que ver con el concepto de número tal como resulta de la operación de *contar*. Es puramente un *símbolo*, sometido a la regla fundamental $i^2 = -1$, y su valor dependerá por completo de si mediante su introducción se ha conseguido una extensión del sistema numérico que resulte útil y manejable.

Puesto que deseamos sumar y multiplicar con el símbolo i en igual forma que con los números reales ordinarios, deberemos ser capaces de formar símbolos tales como $2i$, $3i$, $-i$, $2 + 5i$ o, más en general, $a + bi$, donde a y b son dos números reales cualesquiera. Si estos símbolos han de obedecer a las familiares leyes conmutativa, asociativa y distributiva de la suma y del producto, se tendrá, p. ej.,

$$\begin{aligned}(2 + 3i) + (1 + 4i) &= (2 + 1) + (3 + 4)i = 3 + 7i,\\ (2 + 3i)(1 + 4i) &= 2 + 8i + 3i + 12i^2 =\\ &= (2 - 12) + (8 + 3)i = -10 + 11i.\end{aligned}$$

Guiados por estas consideraciones, comenzamos nuestra exposición sistemática haciendo la siguiente *definición*: llamaremos *número complejo de parte real* a y *parte imaginaria* b a un símbolo de la forma $a + bi$, donde a y b representan dos números reales cualesquiera. Con estos símbolos pueden realizarse las operaciones de adición y multiplicación en igual forma que si i fuera un número real ordinario, salvo que i^2 debe ser sustituído siempre por -1; con mayor precisión, definimos la adición y multiplicación de números complejos mediante las reglas

$$\begin{aligned}(a + bi) + (c + di) &= (a + c) + (b + d)i,\\ (a + bi)(c + di) &= (ac - bd) + (ad + bc)i.\end{aligned} \qquad [3]$$

En particular, se tiene:

$$(a + bi)(a - bi) = a^2 - abi + abi - b^2i^2 = a^2 + b^2. \qquad [4]$$

Mediante estas definiciones se comprueba fácilmente que subsisten para los números complejos las leyes conmutativa, asociativa y distributiva. Además, no solamente la adición y multiplicación, sino también la sustracción y división de dos números complejos, dan lugar a números de la forma $a + bi$, de modo que los números complejos forman un *cuerpo* (véase pág. 64).

$$(a + bi) - (c + di) = (a - c) + (b - d)i,$$

$$\frac{a + bi}{c + di} = \frac{(a + bi)}{(c + di)} \frac{(c - di)}{(c - di)} = \left(\frac{ac + bd}{c^2 + d^2}\right) + \left(\frac{bc - ad}{c^2 + d^2}\right) i. \qquad [5]$$

(La segunda ecuación carece de significado cuando $c + di = 0 + 0i$, pues en este caso $c^2 + d^2 = 0$. Así que de nuevo *debemos excluir la división por cero*; es decir, por $0 + 0i$.) P. ej.:

$$(2 + 3i) - (1 + 4i) = 1 - i,$$

$$\frac{2 + 3i}{1 + 4i} = \frac{2 + 3i}{1 + 4i} \cdot \frac{1 - 4i}{1 - 4i} = \frac{2 - 8i + 3i + 12}{1 + 16} = \frac{14}{17} - \frac{5}{17} i.$$

El cuerpo de los números complejos incluye el cuerpo de los números reales como un subcuerpo de él, pues el número complejo $a + 0i$ puede considerarse igual al número real a. Por otra parte, un número complejo de la forma $0 + bi$ se llama imaginario puro.

Ejercicios:

1. Exprésese $\dfrac{(1 + i)(2 + i)(3 + i)}{(1 - i)}$ en la forma $a + bi$.

2. Exprésese $\left(-\dfrac{1}{2} + i\dfrac{\sqrt{3}}{2}\right)^3$ en la forma $a + bi$.

3. Exprésense en la forma $a + bi$:

$$\frac{1 + i}{1 - i}, \frac{1 + i}{2 - i}, \frac{1}{i^5}, \frac{1}{(-2 + i)(1 - 3i)}, \frac{(4 - 5i)^2}{(2 - 3i)^2}$$

4. Calcúlese $\sqrt{5 + 12i}$. (*Indicación:* Escríbase $\sqrt{5 + 12i} = x + yi$, elévese al cuadrado e iguálense partes reales e imaginarias.)

Mediante la introducción del símbolo i hemos extendido el cuerpo de los números reales al cuerpo de los símbolos $a + bi$, en el cual la ecuación cuadrática especial

$$x^2 = -1$$

tiene las dos soluciones $x = i$ y $x = -i$. En efecto, por definición, $i \cdot i = (-i)(-i) = i^2 = -1$. En realidad, hemos conseguido mucho más, pues es fácil comprobar que ahora *toda ecuación cuadrática*, que podemos escribir en la forma

$$ax^2 + bx + c = 0, \qquad [6]$$

tiene solución. En efecto, de [6] resulta

$$x^2 + \frac{b}{a}x = -\frac{c}{a},$$

$$x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} = \frac{b^2}{4a^2} - \frac{c}{a},$$

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2},$$

$$x + \frac{b}{2a} = \frac{\pm\sqrt{b^2 - 4ac}}{2a},$$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \qquad [7]$$

Ahora bien: si $b^2 - 4ac \geqslant 0$, $\sqrt{b^2 - 4ac}$ es un número real ordinario y las soluciones [7] son reales, mientras que si $b^2 - 4ac < 0$, entonces $4ac - b^2 > 0$ y $\sqrt{b^2 - 4ac} = \sqrt{-(4ac - b^2)} = i\sqrt{4ac - b^2}$, de forma que las soluciones [7] son números complejos; p. ej., las soluciones de la ecuación

$$x^2 - 5x + 6 = 0,$$

son $x = (5 \pm \sqrt{25 - 24})/2 = (5 \pm 1)/2 = 2$ y 3; en tanto que las soluciones de la ecuación

$$x^2 - 2x + 2 = 0,$$

son $x = (2 \pm \sqrt{4 - 8})/2 = (2 \pm 2i)/2 = 1 + i$ y $1 - i$.

2. Interpretación geométrica de los números complejos.—Ya en el siglo XVI los matemáticos se vieron obligados a introducir símbolos para representar las raíces cuadradas de los números negativos y poder resolver todas las ecuaciones cuadráticas y cúbicas. Sin embargo, no fueron capaces de explicar el significado exacto de estos símbolos, que eran considerados con cierto temor supersticioso. El nombre de «imaginarios» es todavía un vestigio del hecho de que estas expresiones fueran consideradas como algo ficticio e irreal. Por fin, a principios del siglo XIX, al ponerse de manifiesto la importancia de estos números en muchas ramas de las matemáticas, una sencilla interpretación geométrica de las operaciones con los números complejos fué suficiente para hacer desaparecer las dudas acerca de su validez. Por supuesto que tal interpretación es innecesaria desde el punto de vista moderno, de acuerdo con el cual la justificación del cálculo formal con los números complejos está basada directamente sobre las definiciones formales de adición y multiplicación. Pero la interpretación geométrica, que

fué dada casi al mismo tiempo por Wessel (1745-1818), Argand (1768-1822) y Gauss, hizo que estas operaciones resultaran más naturales desde un punto de vista intuitivo, habiendo resultado, por otra parte, de la mayor importancia en las aplicaciones de los números complejos a la matemática y a las ciencias físicas.

Esta interpretación geométrica consiste simplemente en representar el número complejo $z = x + yi$ por el punto del plano de coordenadas rectangulares x, y. Así, la parte real de z es su coordenada x, y la parte imaginaria su coordenada y. Con ello queda establecida una correspondencia entre los números complejos y los puntos del «plano numérico», en forma análoga a la correspondencia establecida en II entre los números reales y los puntos de una recta, la recta numérica. Los puntos del eje x del plano numérico corresponden a los números reales $z = x + 0i$, mientras los puntos del eje y corresponden a los números imaginarios puros $z = 0 + yi$.

Si

$$z = x + yi$$

es un número complejo cualquiera, al número complejo

$$\bar{z} = x - yi$$

lo llamaremos *conjugado* de $\bar{z}$. El punto $\bar{z}$ se representa en el plano numérico por el simétrico del punto z respecto al eje x. Si designamos la distancia desde el origen al punto z por ρ, entonces, por el teorema de Pitágoras

$$\rho^2 = x^2 + y^2 = (x + yi)(x - yi) = z \cdot \bar{z}.$$

El número real $\rho = \sqrt{x^2 + y^2}$ se llama *módulo* de z, y se escribe

$$\rho = |z|.$$

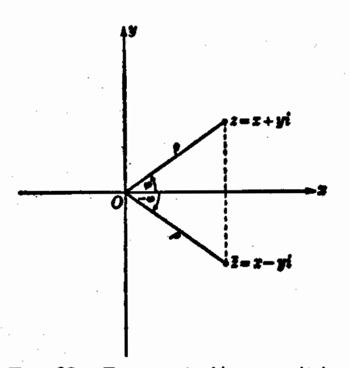


FIG. 22.—Representación geométrica de los números complejos. El punto z tiene como coordenadas rectangulares x e y.

Si z está sobre el eje real, su módulo es su valor absoluto ordinario. Los números complejos de módulo 1 están sobre la «circunferencia unidad», con centro en el origen y radio 1.

Si $|z| = 0$ *entonces* $z = 0$. Esto resulta de la definición de $|z|$ como distancia de z al origen. Además, *el módulo del producto de dos*

números complejos es igual al producto de sus módulos:

$$|z_1 \cdot z_2| = |z_1| \cdot |z_2|.$$

Esta propiedad resulta de un teorema más general que será demostrado en la página 105.

Ejercicios:

1. Demuéstrese este teorema basándose directamente en la definición de producto de dos números complejos $z_1 = x_1 + y_1 i$ y $z_2 = x_2 + y_2 i$.

2. Basándose en el hecho de que el producto de dos números *reales* únicamente es 0 si uno de los factores es 0, demuéstrese el teorema correspondiente para los números *complejos*. (Hágase uso de los dos teoremas que se acaban de enunciar.)

De la definición de suma de dos números complejos, $z_1=x_1+y_1 i$ y $z_2=x_2 + y_2 i$, se tiene

$$z_1 + z_2 = (x_1 + x_2) + (y_1 + y_2)i.$$

Por tanto, el punto $z_1 + z_2$ está representado en el plano numérico por el cuarto vértice de un paralelogramo cuyos otros tres vértices son los puntos 0, z_1, z_2. Esta sencilla construcción geométrica de la suma de dos números complejos es de gran importancia en muchas aplicaciones; de ella puede deducirse la consecuencia importante de que *el módulo de la suma de dos números complejos no excede nunca a la suma de los módulos* (véase la pág. 66):

$$|z_1 + z_2| \leqq |z_1| + |z_2|.$$

FIG. 23.—Construcción geométrica de la suma de dos números complejos.

Esta propiedad resulta del hecho de ser la longitud de un lado cualquiera de un triángulo menor que la suma de las longitudes de los otros dos.

Ejercicio: ¿En qué caso tiene lugar la igualdad $|z_1 + z_2| = |z_1| + |z_2|$?

El ángulo formado por la dirección positiva del eje x y la recta Oz se llama *argumento* de z, y se representa por φ (Fig. 22). El módulo de $\bar{z}$ es el mismo que el de z

$$|\bar{z}| = |z|,$$

pero el argumento de $\bar{z}$ es opuesto al de z; esto es,

$$\bar{\varphi} = -\varphi.$$

Naturalmente, el argumento de z no está determinado unívocamente, ya que puede sumarse o restarse a un ángulo cualquier múltiplo entero de 360°, sin que esto afecte a la posición gráfica de sus lados. Así, p. ej.,

$$\varphi,\ \varphi + 360°,\ \varphi + 720°,\ \varphi + 1080°, \cdots,$$
$$\varphi - 360°,\ \varphi - 720°,\ \varphi - 1080°, \cdots$$

representan todos gráficamente el mismo ángulo. Utilizando el módulo ρ y el argumento φ, el número complejo z puede escribirse en la forma

$$z = x + yi = \rho(\cos\varphi + i \operatorname{sen}\varphi); \qquad [8]$$

en efecto, por definición de seno y coseno (véase Cap. VI, I, 2),

$$x = \rho\cos\varphi, \qquad y = \rho\operatorname{sen}\varphi.$$

Por ejemplo, para $z = i$, $\rho = 1$, $\varphi = 90°$, de manera que $i = 1\,(\cos 90° + i \operatorname{sen} 90°)$;

para $z = 1 + i$, $\rho = \sqrt{2}$, $\varphi = 45°$, de modo que
$$1 + i = \sqrt{2}\,(\cos 45° + i \operatorname{sen} 45°);$$
para $z = 1 - i$, $\rho = \sqrt{2}$, $\varphi = -45°$, de modo que
$$1 - i = \sqrt{2}\,[\cos(-45°) + i \operatorname{sen}(-45°)];$$
para $z = -1 + \sqrt{3}\,i$, $\rho = 2$, $\varphi = 120°$, de forma que
$$-1 + \sqrt{3}\,i = 2\,(\cos 120° + i \operatorname{sen} 120°).$$

El lector puede comprobar estas igualdades por sustitución de los valores de las funciones trigonométricas.

La representación trigonométrica [8] es de gran interés en el caso de la multiplicación de números complejos. Pues si

$$z = \rho(\cos\varphi + i \operatorname{sen}\varphi),$$
y
$$z' = \rho'(\cos\varphi' + i \operatorname{sen}\varphi'),$$
se tiene
$$zz' = \rho\rho'\{(\cos\varphi\cos\varphi' - \operatorname{sen}\varphi\operatorname{sen}\varphi') + i(\cos\varphi\operatorname{sen}\varphi' + \operatorname{sen}\varphi\cos\varphi')\}$$

Ahora bien: por los teoremas de adición del seno y del coseno,

$$\cos\varphi\cos\varphi' - \operatorname{sen}\varphi\operatorname{sen}\varphi' = \cos(\varphi + \varphi'),$$
$$\cos\varphi\operatorname{sen}\varphi' + \operatorname{sen}\varphi\cos\varphi' = \operatorname{sen}(\varphi + \varphi').$$

Por consiguiente,

$$zz' = \rho\rho'\{\cos(\varphi + \varphi') + i \operatorname{sen}(\varphi + \varphi')\}. \qquad [9]$$

Esta es la forma trigonométrica del número complejo de módulo $\rho\rho'$ y argumento $\varphi + \varphi'$. En otras palabras, *para multiplicar dos números complejos se multiplican sus módulos y se suman sus argumentos* (figura 24). Vemos así que la multiplicación de números complejos está relacionada con la operación geométrica de *rotación*. Con más precisión, denominemos al segmento dirigido que une el origen con el punto z *vector* z; la longitud de éste será $\rho = |z|$. Sea z' un número sobre la circunferencia unidad, es decir, $\rho' = 1$; entonces, al multiplicar z por z' el vector z gira un ángulo igual a φ'. Si $\rho' \neq 1$, la longitud del vector ha de ser multiplicada por ρ' una vez efectuada la rotación. El lector puede comprobar gráficamente estos hechos, haciendo el producto de varios números por $z_1 = i$ (giro de 90°); $z_2 = -i$ (giro de 90° en sentido opuesto); $z_3 = 1 + i$, y $z_4 = 1 - i$.

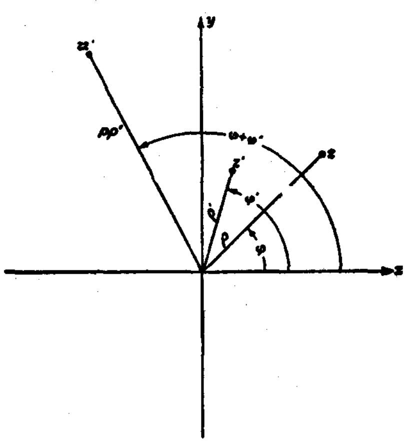


FIG. 24.—Multiplicación de dos números complejos; los argumentos se suman y los módulos se multiplican.

La fórmula [9] tiene una consecuencia particularmente importante cuando $z = z'$, pues en este caso resulta

$$z^2 = \rho^2(\cos 2\varphi + i \operatorname{sen} 2\varphi).$$

Al multiplicar de nuevo por z obtenemos

$$z^3 = \rho^3(\cos 3\varphi + i \operatorname{sen} 3\varphi),$$

y si se continúa indefinidamente en esta forma,

$$z^n = \rho^n(\cos n\varphi + i \operatorname{sen} n\varphi) \qquad \text{para cualquier entero } n. \qquad [10]$$

En particular, si z es un punto de la *circunferencia unidad*, $\rho = 1$, obtenemos la fórmula dada a conocer por el matemático inglés A. De Moivre (1667-1754):

$$(\cos \varphi + i \operatorname{sen} \varphi)^n = \cos n\varphi + i \operatorname{sen} n\varphi. \qquad [11]$$

Esta fórmula es una de las más notables y útiles de las matemáticas elementales, según vamos a poner de manifiesto con un ejemplo.

Si aplicamos la fórmula para $n = 3$ y desarrollamos el primer miembro por la fórmula del binomio

$$(u + v)^3 = u^3 + 3u^2v + 3uv^2 + v^3.$$

se obtiene la relación

$$\cos 3\varphi + i \operatorname{sen} 3\varphi = \cos^3 \varphi - 3 \cos \varphi \operatorname{sen}^2 \varphi + i(3 \cos^2 \varphi \operatorname{sen} \varphi - \operatorname{sen}^3 \varphi).$$

Una sola ecuación, tal como ésta, *entre dos números complejos equivale a un par de ecuaciones entre números reales*, pues la igualdad de dos números complejos exige la igualdad de sus partes reales e imaginarias, respectivamente. En consecuencia podemos escribir

$$\cos 3\varphi = \cos^3 \varphi - 3 \cos \varphi \operatorname{sen}^2 \varphi, \qquad \operatorname{sen} 3\varphi = 3 \cos^2 \varphi \operatorname{sen} \varphi - \operatorname{sen}^3 \varphi.$$

Si utilizamos la igualdad

$$\cos^2 \varphi + \operatorname{sen}^2 \varphi = 1,$$

se obtiene finalmente

$$\begin{aligned} \cos 3\varphi &= \cos^3 \varphi - 3 \cos \varphi(1 - \cos^2 \varphi) = 4 \cos^3 \varphi - 3 \cos \varphi, \\ \operatorname{sen} 3\varphi &= -4 \operatorname{sen}^3 \varphi + 3 \operatorname{sen} \varphi. \end{aligned}$$

Fácilmente pueden obtenerse fórmulas análogas para cualquier valor entero n que expresen sen $n\varphi$ y cos $n\varphi$ en función de las potencias de sen φ y cos φ, respectivamente.

Ejercicios:

1. Hállense las fórmulas correspondientes para sen 4φ y cos 4φ.
2. Demuéstrese que para un punto, $z = \cos \varphi + i \operatorname{sen} \varphi$, de la circunferencia unidad, $1/z = \cos \varphi - i \operatorname{sen} \varphi$.
3. Pruébese, sin efectuar cálculos, que $(a + bi)/(a - bi)$ tiene siempre módulo 1.
4. Si z_1 y z_2 son dos números complejos, demuéstrese que el argumento de $z_1 - z_2$ es igual al ángulo formado por el eje real y el vector dirigido de z_2 a z_1.
5. Interprétese el argumento del número complejo $(z_1 - z_2)/(z_1 - z_3)$ en el triángulo formado por los puntos z_1, z_2 y z_3.
6. Demuéstrese que el cociente de dos números complejos del mismo argumento es real.
7. Demuéstrese que si dados cuatro números complejos z_1, z_2, z_3, z_4, los argumentos de $(z_3 - z_1)/(z_3 - z_2)$ y $(z_4 - z_1)/(z_4 - z_2)$ son iguales, los cuatro puntos están sobre una circunferencia o una recta, y recíprocamente.
8. Demuéstrese que la condición necesaria y suficiente para que los cuatro puntos z_1, z_2, z_3, z_4 estén sobre una circunferencia o una recta es que sea real el cociente

$$\frac{z_3 - z_1}{z_3 - z_2} \Big/ \frac{z_4 - z_1}{z_4 - z_2}$$

3. Fórmula de De Moivre y raíces de la unidad.—Entendemos por raíz n-ésima de un número a aquel número b tal que $b^n = a$. En particular, el número 1 tiene dos raíces cuadradas, 1 y -1, pues $1^2 = (-1)^2 = 1$. El número 1 tiene una sola raíz cúbica real mientras que tiene cuatro raíces cuartas: los números reales 1 y -1 y los imaginarios i y $-i$. Estos hechos sugieren que debe haber otras dos raíces cúbicas de 1 en el campo complejo, haciendo un total de tres. Que efectivamente es así puede verse sin dificultad mediante la fórmula de De Moivre.

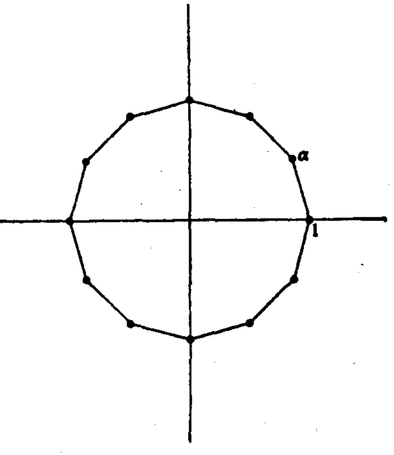


FIG. 25.—Representación geométrica de las raíces dozavas de la unidad.

Vamos a demostrar que *en el cuerpo de los números complejos hay exactamente* n *raíces* n*-ésimas diferentes de 1, las cuales están representadas por los vértices de un* n*-ágono regular inscrito en la circunferencia unidad, siendo uno de sus vértices el punto* z = 1. Esto resulta casi evidente de la figura 25 (dibujada para el caso $n = 12$). El primer vértice del polígono es 1, y el siguiente es

$$\alpha = \cos \frac{360^\circ}{n} + i \operatorname{sen} \frac{360^\circ}{n}, \qquad [12]$$

ya que su argumento debe ser la n-ésima parte del ángulo total de 360°. El vértice inmediato es $\alpha \cdot \alpha = \alpha^2$, pues se obtiene girando el vector α un ángulo de $\frac{360^\circ}{n}$. El vértice siguiente será α^3, etc., y, finalmente, después de reiterar n veces, volvemos nuevamente al vértice 1; esto es, se tiene

$$\alpha^n = 1,$$

lo que también resulta evidente de la fórmula [11], ya que

$$\left[\cos \frac{360^\circ}{n} + i \operatorname{sen} \frac{360^\circ}{n}\right]^n = \cos 360^\circ + i \operatorname{sen} 360^\circ = 1 + 0i.$$

Resulta que $\alpha^1 = \alpha$ es una raíz de la ecuación $x^n = 1$, y lo mismo

ocurre con el vértice inmediato $\alpha^2 = \cos\left(\frac{720^\circ}{n}\right) + i \operatorname{sen}\left(\frac{720^\circ}{n}\right)$
Esto se obtiene sin más que escribir

$$(\alpha^2)^n = \alpha^{2n} = (\alpha^n)^2 = (1)^2 = 1,$$

o, por la fórmula de De Moivre:

$$(\alpha^2)^n = \cos\left(n\frac{720^\circ}{n}\right) + i \operatorname{sen}\left(n\frac{720^\circ}{n}\right) = \cos 720^\circ + i \operatorname{sen} 720^\circ = 1 + 0i = 1.$$

De la misma manera veríamos que todos los n números

$$1,\ \alpha,\ \alpha^2,\ \alpha^3, \cdots,\ \alpha^{n-1}$$

son raíces n-ésimas de 1. Si seguimos adelante en la sucesión de exponentes o utilizamos exponentes negativos, no se obtienen nuevas raíces; en efecto, $\alpha^{-1} = 1/\alpha = \alpha^n/\alpha = \alpha^{n-1}$, y $\alpha^n = 1$, $\alpha^{n+1} = (\alpha)^n \alpha = 1 \cdot \alpha = \alpha$, etc., de forma que se repiten los valores anteriores. Se deja como ejercicio al lector el demostrar que no hay otras raíces n-ésimas.

Si n es par, uno de los vértices del n-ágono coincide con el punto -1, de acuerdo con el hecho algebraico de que en este caso -1 es una raíz n-ésima de 1.

La ecuación a la que satisfacen las raíces n-ésimas de la unidad

$$x^n - 1 = 0 \qquad [13]$$

es de grado n, pero puede reducirse fácilmente a una ecuación de grado $n - 1$, para lo cual basta hacer uso de la identidad algebraica

$$x^n - 1 = (x - 1)(x^{n-1} + x^{n-2} + x^{n-3} + \cdots + 1). \qquad [14]$$

Dado que el producto de dos números es 0 si, *y sólo si*, uno al menos de los factores es 0, el primer miembro de [14] se anula sólo si uno de los dos factores del segundo miembro es cero, esto es, sólo si $x=1$, o bien

$$x^{n-1} + x^{n-2} + x^{n-3} + \cdots + x + 1 = 0. \qquad [15]$$

Por tanto, ésta es la ecuación a la que deben satisfacer las raíces α, $\alpha^2, \ldots, \alpha^{n-1}$, y recibe el nombre de *ecuación ciclotómica* (divisora de la circunferencia); p. ej., las raíces cúbicas imaginarias de 1,

$$\alpha = \cos 120^\circ + i \operatorname{sen} 120^\circ = \tfrac{1}{2}(-1 + i\sqrt{3}),$$
$$\alpha^2 = \cos 240^\circ + i \operatorname{sen} 240^\circ = \tfrac{1}{2}(-1 - i\sqrt{3}),$$

son las raíces de la ecuación

$$x^2 + x + 1 = 0,$$

como puede comprobar el lector por sustitución directa. Análogamente, las raíces quintas de 1, aparte la unidad, satisfacen a la ecuación

$$x^4 + x^3 + x^2 + x + 1 = 0. \qquad [16]$$

Para construir un pentágono regular debemos resolver esta ecuación de cuarto grado, la cual, por un simple artificio algebraico, se reduce a una ecuación cuadrática en $w = x + \frac{1}{x}$. Si dividimos [16] por x^2 y reagrupamos los términos:

$$x^2 + \frac{1}{x^2} + x + \frac{1}{x} + 1 = 0.$$

o bien, ya que $\left(x + \frac{1}{x}\right)^2 = x^2 + \frac{1}{x^2} + 2$, obtenemos la ecuación

$$w^2 + w - 1 = 0.$$

Mediante la fórmula [7] anterior, esta ecuación tiene las raíces

$$w_1 = \frac{-1 + \sqrt{5}}{2}, \qquad w_2 = \frac{-1 - \sqrt{5}}{2}$$

Por consiguiente, las raíces quintas de 1 son las correspondientes a dos ecuaciones cuadráticas

$$x + \frac{1}{x} = w_1, \quad \text{o} \quad x^2 + {}^1/_2(\sqrt{5} - 1)x + 1 = 0,$$

y

$$x + \frac{1}{x} = w_2, \quad \text{o} \quad x^2 - {}^1/_2(\sqrt{5} + 1)x + 1 = 0,$$

que el lector puede resolver por medio de la fórmula ya utilizada.

Ejercicios:

1. Hállense las raíces sextas de 1.
2. Calcúlese $(1 + i)^{11}$.
3. Hállense los diferentes valores de $\sqrt{1 + i}$, $\sqrt[3]{7 - 4i}$, $\sqrt[3]{i}$, $\sqrt[5]{-i}$.
4. Calcúlese $\frac{1}{2i}(i^7 - i^{-7})$.

***4. El teorema fundamental del álgebra.**—No sólo toda ecuación de la forma $ax^2 + bx + c = 0$ ó de la forma $x^n - 1 = 0$ es resoluble en el cuerpo de los números complejos, sino que se verifica en general que *toda ecuación algebraica de grado* n, *de coeficientes reales o complejos,*

$$f(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1x + a_0 = 0, \qquad [17]$$

tiene soluciones en el cuerpo de los números complejos. Este resultado fué establecido para las ecuaciones de tercero y cuarto grados en el siglo XVI por Tartaglia, Cardano y otros, que resolvieron dichas ecuaciones mediante fórmulas esencialmente análogas a la de la ecuación de segundo grado, si bien mucho más complicadas. Durante casi doscientos años se estudiaron con gran insistencia las ecuaciones de quinto grado y grados superiores; pero todos los esfuerzos para resolverlas por métodos similares fracasaron. Fué una gran hazaña del joven Gauss el dar la primera demostración completa en su tesis doctoral (1799) de la *existencia* de soluciones, aunque la cuestión de generalizar las fórmulas clásicas que expresan las soluciones de las ecuaciones de grado inferior al quinto mediante operaciones racionales y extracción de raíces quedó sin respuesta en su tiempo (véase pág. 128).

El teorema de Gauss dice que *para toda ecuación algebraica de la forma* [17], *donde* n *es un entero positivo y los coeficientes números reales cualesquiera o incluso números complejos, existe al menos un número complejo* $\alpha = c + di$ *tal que*

$$f(\alpha) = 0.$$

El número α se llama una *raíz* de la ecuación [17]. Daremos una demostración de este teorema en el Apéndice del capítulo V. Si de momento lo consideramos demostrado, podemos probar el llamado *teorema fundamental del álgebra* (con mayor propiedad debería llamarse teorema fundamental del sistema de los números complejos); esto es, *todo polinomio de grado* n

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0\,, \qquad [18]$$

puede descomponerse en el producto de exactamente n *factores*

$$f(x) = (x - \alpha_1)(x - \alpha_2)\cdots(x - \alpha_n), \qquad [19]$$

siendo $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n$ *números complejos, raíces de la ecuación* $f(x)=0$. Como ejemplo para aclarar este teorema, el polinomio

$$f(x) = x^4 - 1$$

puede escribirse en la forma

$$f(x) = (x - 1)(x - i)(x + i)(x + 1).$$

De la descomposición [19] resulta evidente que las α son raíces de la ecuación $f(x) = 0$, ya que para $x = \alpha_1$, un factor de $f(x)$, y en consecuencia $f(x)$ mismo, es igual a cero.

En algunos casos los factores $(x - \alpha_1)$, $(x - \alpha_2)$, ... de un polinomio $f(x)$ de grado n no son todos distintos, como se ve en el ejemplo

$$f(x) = x^2 - 2x + 1 = (x - 1)(x - 1),$$

que tiene sólo la raíz $x = 1$, «contada dos veces» o «con multiplicidad 2». En todo caso, un polinomio de grado n no puede tener más de n factores distintos $(x - \alpha)$, y la correspondiente ecuación, n raíces.

Para demostrar el teorema de la descomposición en factores, haremos uso de la identidad algebraica

$$x^k - \alpha^k = (x - \alpha)(x^{k-1} + \alpha x^{k-2} + \alpha^2 x^{k-3} + \cdots + \alpha^{k-2}x + \alpha^{k-1}), \quad [20]$$

que para $\alpha = 1$ se reduce a la fórmula de sumación de una progresión geométrica. Como hemos supuesto que se verifica el teorema de Gauss, podemos admitir que $\alpha = \alpha_1$ es una raíz de la ecuación [17], de modo que

$$f(\alpha_1) = \alpha_1^n + a_{n-1}\alpha_1^{n-1} + a_{n-2}\alpha_1^{n-2} + \cdots + a_1\alpha_1 + a_0 = 0.$$

Restando ésta de $f(x)$ y reagrupando los términos, obtenemos la identidad

$$f(x) = f(x) - f(\alpha_1) = (x^n - \alpha_1^n) + a_{n-1}(x^{n-1} - \alpha_1^{n-1}) + \qquad [21]$$
$$+ \cdots + a_1(x - \alpha_1).$$

Ahora bien: según [20], podemos sacar el factor $(x - \alpha_1)$ de cada término de [21], con lo cual el grado del otro factor de cada término se reduce en una unidad. Si de nuevo reagrupamos los términos, se obtiene

$$f(x) = (x - \alpha_1)g(x),$$

donde $g(x)$ es un polinomio de grado $n - 1$:

$$g(x) = x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_1 x + b_0 .$$

(Para nuestro objeto resulta innecesario calcular los coeficientes b_k.)

Ahora podemos aplicar el mismo procedimiento a $g(x)$; por el teorema de Gauss, existe una raíz α_2 de la ecuación $g(x) = 0$, de modo que

$$g(x) = (x - \alpha_2)h(x),$$

siendo $h(x)$ un polinomio de grado $n - 2$. Reiterando el mismo proceso un total de $(n - 1)$ veces (por supuesto que esta frase no es sino un equivalente del proceso de inducción matemática) obtenemos finalmente la descomposición completa en factores

$$f(x) = (x - \alpha_1)(x - \alpha_2)(x - \alpha_3)\dots(x - \alpha_n). \qquad [22]$$

De [22] resulta no sólo que los números complejos $\alpha_1, \alpha_2, \dots, \alpha_n$ son raíces de la ecuación [17], sino también que son sus *únicas* raíces; pues si fuera y una raíz de la ecuación [17], por [22] se tendría

$$f(y) = (y - \alpha_1)(y - \alpha_2)\cdots(y - \alpha_n) = 0.$$

Hemos visto en la página 103 que un producto de números complejos se anula si, *y sólo si*, uno de los factores es igual a 0; en consecuencia, una de las diferencias $(y - \alpha_r)$ debe ser nula, e y igual a α_r, como queríamos demostrar.

*VI. NÚMEROS ALGEBRAICOS Y TRASCENDENTES

1. Definición y existencia.—Un *número algebraico* es cualquier número x, real o complejo, que satisface a una ecuación algebraica de la forma

$$a_nx^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0 \qquad (n \geq 1,\ a_n \neq 0) \qquad [1]$$

donde los coeficientes a_k son *enteros*; p. ej., $\sqrt{2}$ es un número algebraico, pues satisface a la ecuación

$$x^2 - 2 = 0.$$

En forma análoga, toda raíz de una ecuación de coeficientes enteros de grados 3.º, 4.º, 5.º, o superior es un número algebraico, sean o no expresables dichas raíces mediante radicales. El concepto de número algebraico es una generalización natural del de número racional, que constituye el caso especial para $n = 1$.

Que no todo número real es algebraico puede probarse por un procedimiento debido a Cantor, consistente en demostrar que el conjunto de los números algebraicos es *numerable*. Como el conjunto de

todos los números reales no es numerable, de ello resulta la existencia de números reales no algebraicos.

Un método para enumerar el conjunto de los números algebraicos es el siguiente: a cada ecuación de la forma [1] se le adjunta el entero positivo

$$h = |a_n| + |a_{n-1}| + \cdots + |a_1| + |a_0| + n$$

que llamaremos su *altura*. Para un valor *fijo* de h existe sólo un número *finito* de ecuaciones [1] de altura h, y cada una de éstas tiene a lo sumo n raíces diferentes; por tanto, existe sólo un número finito de números algebraicos cuyas ecuaciones son de altura h, y podemos ordenar todos los números algebraicos en una sucesión, partiendo de los de altura 1, a continuación los de altura 2, y así sucesivamente.

Esta demostración de que el conjunto de los números algebraicos es numerable asegura la existencia de números no algebraicos; tales números se llaman *trascendentes*, pues, como dijo Euler, «trascienden al poder de los métodos algebraicos».

La demostración de Cantor de la existencia de números trascendentes a duras penas puede llamarse constructiva. Teóricamente, se puede construir un número trascendente aplicando el proceso diagonal de Cantor a un cuadro numerable de las expresiones decimales de las raíces de las ecuaciones algebraicas; pero este procedimiento es por completo impracticable y no conduciría a ningún número del que se pudiera escribir realmente su expresión en el sistema decimal o en otro cualquiera. Por otra parte, los problemas más interesantes relativos a los números trascendentes se reducen a demostrar que ciertos y determinados números, tales como π y e (estos números serán definidos en el capítulo IV) son, en efecto, trascendentes.

****2. El teorema de Liouville y la construcción de números trascendentes.**—Una demostración de la existencia de números trascendentes anterior a la de Cantor fué dada por J. Liouville (1809-1882), y permite efectivamente *construir* ejemplos de tales números. Es algo más difícil que la demostración de Cantor, como ocurre casi siempre con las demostraciones constructivas comparadas con las que se limitan a probar la existencia. Incluímos la demostración para conocimiento del lector más preparado, aunque tampoco requiere gran bagaje matemático.

Liouville demostró que los números algebraicos irracionales son aquellos que no pueden ser aproximados mediante números racionales con un alto grado de exactitud, salvo que los denominadores de las fracciones que constituyen dichas aproximaciones sean muy grandes.

Supongamos que el número z satisface a la ecuación algebraica de coeficientes enteros

$$f(x) = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n = 0 \qquad (a_n \neq 0), \qquad [2]$$

pero no a una ecuación de grado inferior. Se dice entonces que z es un número algebraico *de grado n*; p. ej., $z = \sqrt{2}$ es un número algebraico de grado 2 por satisfacer a la ecuación $x^2 - 2 = 0$ y no ser raíz de ninguna ecuación de primer grado; $z = \sqrt[3]{2}$ es de tercer grado, ya que satisface a la ecuación $x^3 - 2 = 0$ y, según hemos visto en el capítulo III, a ninguna ecuación de grado inferior. Un número algebraico de grado $n > 1$ no puede ser racional, pues un número racional $z=p/q$ satisface a la ecuación de grado 1, $qx - p = 0$. Ahora bien: todo número irracional z puede aproximarse hasta el grado de exactitud deseado por un número racional, lo que significa que podemos determinar una sucesión

$$\frac{p_1}{q_1}, \frac{p_2}{q_2}, \ldots$$

de números racionales con denominadores crecientes tales que

$$\frac{p_r}{q_r} \to z.$$

El teorema de Liouville afirma que para todo número algebraico z de grado $n > 1$ tal aproximación debe adolecer de un error mayor que $1/q^{n+1}$; esto es, que la desigualdad

$$\left| z - \frac{p}{q} \right| > \frac{1}{q^{n+1}} \qquad [3]$$

subsiste para denominadores q suficientemente grandes.

Nos proponemos demostrar el teorema; pero en primer lugar vamos a ver cómo permite la construcción efectiva de números trascendentes. Sea el número (véase pág. 25 para la definición del símbolo $n!$)

$$z = a_1 \cdot 10^{-1!} + a_2 \cdot 10^{-2!} + a_3 \cdot 10^{-3!} + \cdots + a_m \cdot 10^{-m!} +$$
$$+ a_{m+1} \cdot 10^{-(m+1)!} + \cdots = 0{,}a_1a_2000a_3000000000000000000a_40000000\ldots.$$

donde los a_i son números dígitos arbitrarios de 1 a 9 (podemos, p. ej., elegir todos los a_i iguales a 1). Tal número se caracteriza por el rápido crecimiento de las series de ceros, interrumpidas por dígitos aislados

no ceros. Designemos por z_m la fracción decimal finita obtenida tomando sólo los términos de z hasta incluir el $a_m \cdot 10^{-m!}$. Entonces

$$|z - z_m| < 10 \cdot 10^{-(m+1)!}. \qquad [4]$$

Supongamos que z fuera algebraico de grado n, y pongamos en [3] $p/q = z_m = p/10^{m!}$, con lo que obtenemos

$$|z - z_m| > \frac{1}{10^{(n+1)m!}}$$

para m suficientemente grande. Combinando esta desigualdad con la [4] tendríamos

$$\frac{1}{10^{(n+1)m!}} < \frac{10}{10^{(m+1)!}} = \frac{1}{10^{(m+1)!-1}},$$

de modo que $(n + 1)m! > (m + 1)! - 1$ para todo m suficientemente grande, y esto es evidentemente falso para todo valor de m mayor que n (el lector puede hacer una demostración detallada de esta afirmación), lo que conduce a una contradicción; en consecuencia, z es trascendente.

Queda ahora por demostrar el teorema de Liouville. Supongamos que z es un número algebraico de grado $n > 1$ que satisface a [1], de forma que

$$f(z) = 0. \qquad [5]$$

Sea $z_m = p_m/q_m$ una sucesión de números racionales tales que $z_m \to z$. Entonces

$$f(z_m) = f(z_m) - f(z) = a_1(z_m - z) + a_2(z_m^2 - z^2) + \cdots + a^n(z_m^n - z^n).$$

Dividiendo ambos miembros de la ecuación por $z_m - z$, y utilizando la identidad algebraica

$$\frac{u^n - v^n}{u - v} = u^{n-1} + u^{n-2}v + u^{n-3}v^2 + \cdots + uv^{n-2} + v^{n-1}$$

obtenemos

$$\frac{f(z_m)}{z_m - z} = a_1 + a_2(z_m + z) + a_3(z_m^2 + z_m z + z^2) + \cdots + \qquad [6]$$
$$+ a_n(z_m^{n-1} + \cdots + z^{n-1}).$$

Como z_m tiene por límite z, para valores grandes de m diferirá de z en menos de 1, por lo que podemos escribir, para valores suficientemente grandes de m,

$$\left|\frac{f(z_m)}{z_m - z}\right| < |a_1| + 2|a_2|(|z| + 1) + 3|a_3|(|z| + 1)^2 + \cdots + \qquad [7]$$
$$+ n|a_n|(|z| + 1)^{n-1} = M,$$

que es un número fijo, ya que suponemos fijo z en nuestro razonamiento. Si ahora elegimos m suficientemente grande para que el denominador q_m de $z_m = p_m/q_m$ sea mayor que M, se tendrá

$$|z - z_m| > \frac{|f(z_m)|}{M} > \frac{|f(z_m)|}{q_m} \qquad [8]$$

Por brevedad escribamos p y q en lugar de p_m y q_m: Entonces

$$|f(z_m)| = \left|\frac{a_0q^n + a_1q^{n-1}p + \cdots + a_np^n}{q^n}\right| \qquad [9]$$

Ahora bien: el número racional $z_m = p/q$ no puede ser raíz de $f(x)=0$, pues si así fuese se podría suprimir el factor $(x - z_m)$ de $f(x)$, y z satisfaría a una ecuación de grado inferior a n; por consiguiente, $f(z_m) \neq 0$. Pero el numerador del segundo miembro de [9] es entero; por lo menos, pues, igual a 1. Finalmente, de [8] y [9] tenemos

$$|z - z_m| > \frac{1}{q}\frac{1}{q^n} = \frac{1}{q^{n+1}}, \qquad [10]$$

lo que demuestra el teorema

Durante las últimas décadas ha avanzado mucho la investigación acerca de la posibilidad de aproximar los números algebraicos mediante los racionales; p. ej., el matemático noruego A. Thue (1863-1922) demostró que en la desigualdad [3] de Liouville el exponente $n + 1$ puede ser sustituído por $(n/2) + 1$. C. L. Siegel probó posteriormente el enunciado aún más preciso (más preciso para valores grandes de n) de que subsiste para el exponente $2\sqrt{n}$.

El tema de los números trascendentes ha fascinado siempre a los matemáticos, pero hasta época muy reciente se conocían muy pocos ejemplos de números de interés intrínseco de los que se supiera eran trascendentes (en el capítulo III discutiremos el carácter trascendente de π, del cual resulta la imposibilidad de efectuar la cuadratura del círculo con la regla y el compás). En una famosa comunicación al Congreso internacional de matemáticas celebrado en París en 1900, David Hilbert propuso treinta problemas matemáticos fáciles de formular, algunos incluso en lenguaje elemental y hasta popular, pero ninguno de ellos resuelto y ni siquiera inmediatamente accesible a la

técnica matemática entonces existente. Estos «problemas de Hilbert» fueron un reto al subsiguiente período del desarrollo matemático, y casi todos han sido ya resueltos, constituyendo a menudo su solución un claro progreso de la potencia del instrumento matemático y de sus métodos generales. Uno de los problemas que parecían más inaccesibles consistía en demostrar la trascendencia del número

$$2^{\sqrt{2}}$$

o al menos probar que se trataba de un número irracional; durante más de treinta años no surgió ni la más remota esperanza de hallar un método prometedor de atacar el problema. Por fin Siegel e, independientemente, el joven matemático ruso A. Gelfond, descubrieron nuevos métodos para demostrar el carácter trascendente de muchos números importantes de la matemática, incluído el número de Hilbert $2^{\sqrt{2}}$, y más en general, cualquier número de la forma a^b siendo a un número algebraico $\neq 0$ ó 1, y b un número algebraico irracional.

SUPLEMENTO AL CAPÍTULO II

EL ÁLGEBRA DE LOS CONJUNTOS

1. Teoría general.—El concepto de *clase* o *conjunto* de objetos es uno de los más fundamentales de la matemática. Se define un conjunto mediante una propiedad o atributo $\mathfrak{A}$ que cada objeto considerado debe poseer o no; aquellos objetos que poseen la propiedad forman un conjunto correspondiente A. Así, si consideramos los enteros, y la propiedad $\mathfrak{A}$ es la de ser primo, el conjunto correspondiente A es el constituído por todos los números primos 2, 3, 5, 7, ...

El estudio matemático de los conjuntos se basa en el hecho de que éstos pueden ser combinados mediante ciertas operaciones para formar otros conjuntos, al igual que los números se combinan por adición y multiplicación para dar lugar a otros números. El estudio de las operaciones con los conjuntos constituye el «álgebra de conjuntos», que tiene muchas semejanzas formales (aunque también presenta diferencias) con el álgebra de los números. El hecho de que los métodos algebraicos puedan aplicarse al estudio de objetos no numéricos, como los conjuntos, pone de manifiesto la gran generalidad de conceptos de la matemática moderna. En los últimos años se ha visto que el álgebra de los conjuntos ilumina muchas ramas de la matemática, tales como la teoría de la medida y la teoría de las probabilidades; resulta también valiosa en la reducción sistemática de los conceptos matemáticos a sus fundamentos lógicos.

En lo que sigue I representará un conjunto fijo de objetos de naturaleza cualquiera que llamaremos conjunto universal o universo, y A, B, C, ... representarán subconjuntos arbitrarios de I. Si I es el conjunto de todos los enteros, A puede representar el conjunto de los enteros pares, B el de los impares, C el de los números primos, etc. O bien I puede ser el conjunto de todos los puntos de un plano dado, A el de todos los puntos interiores a una circunferencia determinada, B el conjunto de todos los puntos de otro círculo del plano, etc. Por conveniencia incluímos como «subconjuntos» de I al propio conjunto I y al «conjunto vacío» O, que no contiene elementos. El propósito de esta generalización artificial es el de conservar la regla de que a toda propiedad $\mathfrak{A}$ corresponda el subconjunto A de todos los elementos de I que poseen dicha propiedad. En el caso de que $\mathfrak{A}$ sea alguna pro-

piedad válida universalmente, tal como la expresada por la ecuación trivial $x = x$, el correspondiente subconjunto de I será el mismo I, puesto que todo objeto satisface a esta ecuación, en tanto que si $\mathfrak{A}$ es alguna propiedad contradictoria en sí misma, como $x \neq x$, el correspondiente subconjunto no contendrá ningún objeto y lo representaremos por el símbolo O.

El conjunto A se dice que es un *subconjunto* del B si no hay ningún objeto en A que no esté también en B. En este caso escribimos

$$A \subset B \quad \text{o} \quad B \supset A.$$

P. ej., el conjunto A de todos los enteros múltiplos de 10 es un subconjunto del conjunto B de los enteros múltiplos de 5, ya que todo múltiplo de 10 lo es también de 5. La afirmación de $A \subset B$ no excluye la posibilidad de que sea $B \subset A$. Si ambas relaciones tienen lugar, decimos que los conjuntos A y B son iguales y escribimos

$$A = B.$$

Para que esto se verifique, cada elemento de A debe pertenecer a B, y recíprocamente; de modo que los conjuntos A y B contienen exactamente los mismos elementos.

La relación $A \subset B$ tiene muchas analogías con la relación de orden $a \leqslant b$ entre los números reales. En particular se verifica que

1) $A \subset A$.
2) Si $A \subset B$ y $B \subset A$, $A = B$.
3) Si $A \subset B$ y $B \subset C$, $A \subset C$.

Por esta razón la relación $A \subset B$ se denomina una «relación de orden». Su principal diferencia con la relación $a \leqslant b$ para los números consiste en que, mientras para *cada* par de números a y b subsiste siempre una de las relaciones $a \leqslant b$ ó $b \leqslant a$, esto no se verifica para los conjuntos; p. ej., si A designa el conjunto formado por los enteros 1, 2, 3,

$$A = \{1, 2, 3\},$$

y B el formado por los números 2, 3, 4,

$$B = \{2, 3, 4\},$$

entonces ni $A \subset B$ ni $B \subset A$. Por esta razón, la relación $A \subset B$ se dice que determina una ordenación parcial de los conjuntos, mientras la relación $a \leqslant b$ determina una ordenación completa entre los números.

De pasada observemos que en virtud de la definición de la relación $A \subset B$ resulta que

4) $O \subset A$ para cualquier conjunto A, y
5) $A \subset I$,

siendo A cualquier subconjunto del universo I. La relación 4) puede resultar un tanto paradójica, pero se halla de acuerdo con la interpretación estricta del signo $\subset$. Pues la afirmación $O \subset A$ sería falsa sólo si el conjunto vacío O contuviese un objeto no contenido en A, y como el conjunto vacío no contiene objeto alguno, resulta esto imposible, cualquiera que sea el conjunto A.

Vamos a definir ahora dos operaciones con conjuntos que tienen muchas de las propiedades algebraicas de la adición y multiplicación ordinarias de números, aunque conceptualmente son muy distintas. Sean A y B dos conjuntos cualesquiera; entenderemos por «unión» o «suma lógica» de A y B el conjunto formado por todos los objetos que pertenecen, bien a A o a B (incluídos los que puedan pertenecer a ambos), y este conjunto lo representaremos por el símbolo $A + B$. Definimos la «intersección» o «producto lógico» de A y B como el conjunto formado únicamente por aquellos elementos que pertenecen a *ambos* conjuntos A y B, y lo representaremos por el símbolo $A \cdot B$, o simplemente AB. Para aclarar estas operaciones, sean de nuevo A y B los conjuntos

$$A = \{1, 2, 3\}, \qquad B = \{2, 3, 4\},$$

Entonces $$A + B = \{1, 2, 3, 4\}, \qquad AB = \{2, 3\}.$$

Entre las propiedades algebraicas importantes de las operaciones $A + B$ y AB se encuentran las siguientes, que pueden ser comprobadas por el lector basándose en la definición de las mismas:

6) $A + B = B + A$ 7) $AB = BA$
8) $A + (B + C) = (A + B) + C$ 9) $A(BC) = (AB)C$
10) $A + A = A$ 11) $AA = A$
12) $A(B + C) = (AB + AC)$ 13) $A + (BC) = (A + B)(A + C)$
14) $A + O = A$ 15) $AI = A$
16) $A + I = I$ 17) $AO = O$
18) la relación $A \subset B$ es equivalente a una de las dos relaciones $A + B = B$, $AB = A$.

La verificación de estas leyes es un problema de lógica elemental; p. ej., 10) dice que el conjunto formado por aquellos objetos que pertenecen, bien a A o a A es precisamente el conjunto A, mientras 12) afirma que el conjunto formado por aquellos objetos que son de A

y también de B o de C es el mismo que el formado por aquellos objetos que pertenecen, bien a ambos A y B, o a ambos A y C. El razonamiento lógico implicado en este y otros razonamientos puede aclararse mediante la representación de los conjuntos A, B, C por recintos planos, teniendo cuidado de considerar todas las posibilidades en cuanto a que los conjuntos en cuestión tengan elementos distintos y comunes entre sí.

Habrá observado el lector que las leyes 6), 7), 8), 9) y 12) coinciden con las familiares leyes conmutativa, asociativa y distributiva del álgebra, de donde resulta que todas las reglas del álgebra ordinaria de números que sean consecuencia de dichas leyes son también váli-

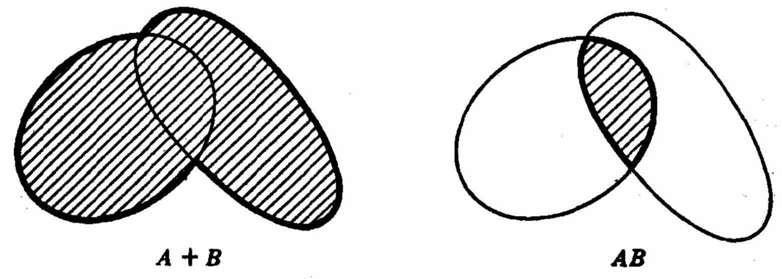


FIG. 26.—Adición e intersección de dos conjuntos.

das en el álgebra de los conjuntos. Las leyes 10), 11) y 13), por otra parte, no tienen análogas en el caso de los números, lo que da al álgebra de conjuntos una estructura más sencilla que la del álgebra de los números; p. ej., el teorema del binomio del álgebra ordinaria, en el álgebra de los conjuntos viene reemplazado por la igualdad

$$(A + B)^n = (A + B) \cdot (A + B) \cdot \cdots \cdot (A + B) = A + B,$$

que es consecuencia de 11). Las leyes 14), 15) y 17) nos dicen que las propiedades de O e I respecto a la unión e intersección de conjuntos son muy similares a las propiedades de los números 0 y 1 respecto a la adición y multiplicación ordinarias. La ley 16) no tiene análoga en el álgebra de números.

Queda por definir una ulterior operación del álgebra de los conjuntos: sea A un subconjunto cualquiera del conjunto universal I. Por *complemento* de A en I entendemos el conjunto formado por todos los objetos de I que *no* pertenecen a A, y representaremos este conjunto por el símbolo A'. Así, si I es el conjunto de todos los números naturales y A el de los números primos, A' estará formado por 1 y todos los números compuestos. La operación A', que no tiene analogía

exacta en el álgebra de los números, goza de las siguientes propiedades:

19) $A + A' = I$ 20) $AA' = O$
21) $O' = I$ 22) $I' = O$
23) $A'' = A$
24) La relación $A \subset B$ es equivalente a la relación $B' \subset A'$.
25) $(A + B)' = A'B'$ 26) $(AB)' = A' + B'$.

Dejamos de nuevo a cargo del lector la verificación de estas leyes.

Las leyes 1) a 26) constituyen la base del álgebra de los conjuntos, y poseen la notable propiedad de «dualidad», entendida en el sentido siguiente: *Si en una cualquiera de las leyes* 1) *a* 26) *los símbolos*

$$\begin{array}{ccc} \subset & \text{y} & \supset \\ O & \text{e} & I \\ + & \text{y} & \cdot \end{array}$$

se intercambian (dondequiera que aparezcan), el resultado es de nuevo una de dichas leyes.

P. ej., la ley 6) se transforma en la 7); la 12) en la 13); la 17) en la 16), etc. De ello resulta que *a todo teorema que pueda demostrarse basándose en las leyes* 1) *a* 26) *corresponde otro teorema «dual», que se obtiene mediante los intercambios indicados*. En efecto, como la demostración de todo teorema consiste en la aplicación sucesiva en cada etapa de alguna de las leyes 1) a 26), la aplicación a cada etapa de la ley dual nos dará la demostración del teorema dual (para una dualidad análoga en geometría, véase Cap. IV).

2. Aplicación a la lógica matemática.—La verificación de las leyes del álgebra de conjuntos se apoya en el análisis del significado lógico de la relación $A \subset B$ y de las operaciones $A + B$, AB y A'. Podemos ahora invertir este proceso y utilizar las leyes 1) a 26) como fundamento de un «álgebra de la lógica». Con más precisión, aquella parte de la lógica concerniente a los conjuntos, o lo que es equivalente, las propiedades o *atributos* de los objetos, puede reducirse a un sistema algebraico formal basado en las leyes 1) a 26). El «universo lógico» define el conjunto I; *cada propiedad o atributo* $\mathfrak{A}$ *de los objetos define el conjunto* A, *constituido por todos los objetos de* I *que poseen este atributo.* Las reglas para traducir la terminología lógica usual al lenguaje de los conjuntos quedan aclaradas con los siguientes ejemplos:

«Bien A o B»	$A + B$
«Ambos A y B»	AB
«No A»	A'

«Ni A ni B»	$(A + B)'$, o de forma equivalente $A'B'$
«No ambos A y B»	$(AB)'$, o de forma equivalente $A' + B'$
«Todo A es B» o «Si A también B» o «A implica B»	$A \subset B$
«Algunos A son B»	$AB \neq O$
«Ningún A es B»	$AB = O$
«Algunos A no son B»	$AB' \neq O$
«No hay ningún A»	$A = O$

En términos del álgebra de conjuntos, el silogismo «Barbara», que dice: «Si todo A es B, y todo B es C, entonces todo A es C», se escribe sencillamente

3) Si $A \subset B$ y $B \subset C$, entonces $A \subset C$.

Análogamente, la «ley de contradicción», que dice: «Un objeto no puede poseer simultáneamente un atributo y no poseerlo», se escribe

20) $AA' = O$,

mientras la «ley del *tertio excluso*» que dice: «un objeto debe poseer un atributo o no poseerlo» se transforma en

19) $A + A' = I$.

Así, pues, la parte de la lógica que puede expresarse con los símbolos $\subset$, $+$, $\cdot$, y $'$ puede tratarse como un sistema algebraico formal sometido a las leyes 1) a 26). Esta fusión del análisis lógico de las matemáticas y del análisis matemático de la lógica ha dado lugar a una nueva disciplina, la *lógica matemática*, que se halla actualmente en vías de vigoroso desarrollo.

Desde el punto de vista de la axiomática es notable el hecho de que las proposiciones 1) a 26), junto con todos los teoremas del álgebra de conjuntos, puedan deducirse de las tres ecuaciones siguientes:

27)
$$\begin{aligned} A + B &= B + A \\ (A + B) + C &= A + (B + C) \\ (A' + B')' + (A' + B)' &= A. \end{aligned}$$

Se deduce de ello que el álgebra de conjuntos puede construirse como una teoría puramente deductiva, al igual que la geometría euclídea, sobre la base de estas tres proposiciones aceptadas como axiomas. Cuando se ha hecho esto, la operación AB y la relación de orden $A \subset B$ se *definen* a partir de $A + B$ y A':

AB representa el conjunto $(A' + B')'$
$A \subset B$ significa que $A + B = B$.

Un ejemplo completamente distinto de un sistema matemático que satisface a todas las leyes formales del álgebra de conjuntos nos lo procuran los ocho números 1, 2, 3, 5, 6, 10, 15, 30, donde $a + b$ se define como el mínimo común múltiplo de a y b; ab, como el máximo común divisor de a y b; $a \subset b$, como la afirmación «a es un factor de b», y a', como el número $30/a$. La existencia de ejemplos tales ha llevado al estudio de los sistemas algebraicos generales que satisfacen a las leyes 27). Dichos sistemas se llaman «álgebras de Boole» en honor de George Boole (1815-1864), matemático y lógico inglés, autor del libro *An Investigation of the Laws of Thought*, publicado en 1854.

3. Una aplicación a la teoría de las probabilidades.—El álgebra de conjuntos aclara notablemente la teoría de las probabilidades. Para considerar sólo el caso más sencillo, imaginemos un experimento con un número finito de resultados posibles, todos los cuales se supondrán «igualmente probables». El experimento puede consistir, p. ej., en sacar una carta al azar de una baraja de 52 cartas perfectamente barajadas. Si el conjunto de los resultados posibles del experimento se representa por I, y si A designa cualquier subconjunto de I, entonces la probabilidad de que el resultado del experimento pertenezca al subconjunto A se define por el cociente

$$p(A) = \frac{\text{número de elementos de } A}{\text{número de elementos de } I}$$

Si designamos el número de elementos de un conjunto A por el símbolo $n(A)$, esta definición puede escribirse en la forma

$$p(A) = \frac{n(A)}{n(I)} \qquad [1]$$

En nuestro ejemplo, si A representa el subconjunto de «corazones», entonces $n(A) = 13$, $n(I) = 52$ y $p(A) = 13/52 = 1/4$.

Los conceptos del álgebra de conjuntos intervienen en el cálculo de probabilidades cuando se conocen las probabilidades de ciertos conjuntos y se desean calcular las correspondientes a otros; p. ej., si conocemos $p(A)$, $p(B)$ y $p(AB)$, podemos calcular la probabilidad $p(A + B)$:

$$p(A + B) = p(A) + p(B) - p(AB). \qquad [2]$$

La demostración es inmediata; en efecto,

$$n(A + B) = n(A) + n(B) - n(AB),$$

ya que los elementos comunes a A y a B, esto es, los elementos de AB, están contados dos veces en la suma $n(A) + n(B)$, y, en consecuencia, debemos restar de esta suma $n(AB)$ para obtener el resultado correcto que corresponde a $n(A + B)$. Dividiendo cada término de esta igualdad por $n(I)$ obtenemos la ecuación [2].

Se obtiene una fórmula más interesante al considerar tres subconjuntos, A, B, C, de I. De [2] resulta

$$p(A + B + C) = p[(A + B) + C] = p(A + B) + p(C) - p[(A + B)C].$$

Por la ley 12) de la página 120 sabemos que $(A+B)C = AC + BC$. En consecuencia,

$$p[(A+B)C] = p(AC+BC) = p(AC) + p(BC) - p(ABC).$$

Sustituyendo en la ecuación previa este valor de $p[(A+B)C]$ y el valor de $p(A+B)$ dado por [2], obtenemos el resultado deseado:

$$p(A+B+C) = p(A) + p(B) + \\ + p(C) - p(AB) - p(AC) - p(BC) + p(ABC). \qquad [3]$$

Como ejemplo, consideremos el siguiente experimento: se escriben al azar los tres dígitos 1, 2, 3 y se pide la probabilidad de que uno al menos ocupe su correspondiente lugar. Sea A el conjunto de todas las permutaciones en las que el 1 está en primer lugar; B el de aquellas donde el 2 está en segundo lugar, y C el conjunto de todas las permutaciones en las que el 3 está en tercer lugar. Deseamos, por tanto, calcular $p(A+B+C)$. Es obvio que

$$p(A) = p(B) = p(C) = \frac{2}{6} = \frac{1}{3};$$

pues cuando un dígito ocupa su propio lugar hay dos ordenaciones posibles para los restantes dentro del total de $3 \cdot 2 \cdot 1 = 6$ permutaciones posibles de los tres dígitos. Además,

$$p(AB) = p(AC) = p(BC) = \frac{1}{6}$$

y

$$p(ABC) = \frac{1}{6},$$

ya que hay sólo una forma de presentarse para cada uno de estos casos. De [3] resulta que

$$p(A+B+C) = 3 \cdot \frac{1}{3} - 3\left(\frac{1}{6}\right) + \frac{1}{6} =$$

$$= 1 - \frac{1}{2} + \frac{1}{6} = \frac{2}{3} = 0{,}6666 \cdots$$

Ejercicio: Hállese la fórmula correspondiente a $p(A+B+C+D)$ y aplíquese al caso de cuatro dígitos. La probabilidad correspondiente es $5/8 = 0{,}6250$.

La fórmula general para la suma de n subconjuntos es

$$p(A_1 + A_2 + \cdots + A_n) = \sum_1 p(A_i) - \sum_2 p(A_iA_j) + \sum_3 p(A_iA_jA_k) - \\ - \cdots \pm p(A_1A_2 \cdots A_n), \qquad [4]$$

donde los símbolos $\sum_1, \sum_2, \sum_3, \cdots, \sum_{n-1}$ representan las sumas de todas las combinaciones posibles de los conjuntos $A_1, A_2, \cdots, A_n$ tomados uno a uno, dos a dos, $\cdots$, $(n-1)$ a $(n-1)$. Esta fórmula puede establecerse por inducción matemática de

igual modo que dedujimos [3] a partir de [2]. De [4] es fácil probar que si los n dígitos 1, 2, 3, · · · , n se escriben al azar, la probabilidad de que al menos un dígito ocupe su propio lugar es

$$p_n = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \cdots \pm \frac{1}{n!}, \qquad [5]$$

en la cual el último término se toma con signo más o menos, según sea n par o impar. En particular, para $n = 5$, la probabilidad es

$$p_5 = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \frac{1}{5!} = \frac{19}{30} = 0{,}63333\cdots.$$

Veremos en el capítulo VIII que al tender n a infinito, la expresión

$$S_n = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \cdots \pm \frac{1}{n!}$$

tiende al límite $1/e$, cuyo valor con cinco decimales es 0,36788. Dado que, por [5], $p_n = 1 - S_n$, queda demostrado que al tender n a infinito,

$$p_n \to 1 - 1/e = 0{,}63212.$$

CAPÍTULO III

CONSTRUCCIONES GEOMÉTRICAS ÁLGEBRA DE LOS CUERPOS NUMÉRICOS

Introducción.—Los problemas de construcción han sido siempre un tema favorito en geometría. Sólo con el auxilio de la regla y el compás puede realizarse una gran variedad de construcciones, como el lector debe recordar; puede construirse el punto medio de un segmento o la bisectriz de un ángulo; puede trazarse una recta perpendicular a otra desde un punto dado; puede ser inscrito un hexágono regular en una circunferencia, etc. En todos estos problemas, la regla se usa meramente como borde rectilíneo para dibujar rectas, pero no para medir o transportar distancias. La restricción tradicional de utilizar como únicos instrumentos la regla y el compás se remonta a la antigüedad, aunque los griegos no vacilaban en usar otros instrumentos.

Uno de los más famosos entre los problemas clásicos de construcción es el llamado problema de Apolonio (hacia el año 200 a. de J.C.), en el cual se dan tres circunferencias arbitrarias del plano y se pide trazar una cuarta, tangente a las tres. En particular, una o más de las circunferencias dadas puede degenerar en un punto o en una recta («circunferencia» de radio «cero» o «infinito», respectivamente); p. ej., se puede construir una circunferencia tangente a dos rectas dadas y que pase por un punto dado. Mientras resulta fácil resolver dichos casos especiales, el problema general es considerablemente más difícil.

Entre todos los problemas de construcción, el de trazar con regla y compás el polígono regular de n lados tiene quizá el máximo interés. Para ciertos valores de n, p. ej., $n = 3, 4, 5, 6$, la solución se conoce desde la antigüedad, y forma parte importante de la geometría elemental. Pero para el heptágono regular ($n = 7$) se ha demostrado que la construcción es imposible. Hay otros tres problemas griegos clásicos para los cuales se ha buscado en vano una solución: la trisección de un ángulo arbitrario, la duplicación del cubo (es decir, la construcción de la arista de un cubo cuyo volumen sea doble del de un cubo de arista dada) y la cuadratura del círculo (esto es, la construcción de un cuadrado que tenga igual área que un círculo dado). En todos estos problemas los únicos instrumentos permitidos son la regla y el compás. Problemas de este tipo, que seguían sin re-

solverse, dieron nacimiento a uno de los más notables desarrollos de la matemática, cuando, después de siglos de investigación inútil, surgió la sospecha de que dichos problemas debían de ser definitivamente irresolubles. Los matemáticos se propusieron investigar la cuestión de: *¿Cómo es posible probar que ciertos problemas no pueden resolverse?*

En álgebra, fué el problema de resolver ecuaciones de quinto a superior grado el que llevó a esta nueva manera de pensar. Durante el siglo XVI los matemáticos habían aprendido que las ecuaciones algebraicas de tercero o cuarto grados podían resolverse por un proceso análogo al método elemental utilizado para resolver las ecuaciones cuadráticas. Todos estos métodos tienen la siguiente característica común: las soluciones o «raíces» de la ecuación pueden escribirse como expresiones algebraicas, a partir de los coeficientes de aquélla, mediante una sucesión de operaciones, cada una de las cuales es una operación racional —adición, resta, multiplicación o división— o la extracción de una raíz cuadrada, cúbica o cuarta. Se dice que las ecuaciones algebraicas hasta las de cuarto grado pueden ser resueltas «por radicales» (*radix* es la palabra latina de raíz). Nada parece más natural que extender este procedimiento a las ecuaciones de grados quinto o superior, utilizando raíces de índice mayor, pero tales intentos fracasaron; incluso algunos distinguidos matemáticos del siglo XVIII se engañaron creyendo que habían dado con la solución. Ya en los primeros años del siglo XIX el italiano Ruffini (1765-1822) y el genio noruego N. H. Abel (1802-1829) concibieron la entonces revolucionaria idea de probar *la imposibilidad de resolver la ecuación algebraica general de grado* n *por medio de radicales.*

Debe quedar completamente claro que la cuestión no reside en si cualquier ecuación algebraica de grado n posee soluciones; este hecho fué demostrado por vez primera por Gauss en su tesis doctoral en 1799. Así, pues, no hay duda acerca de la *existencia* de soluciones de una ecuación, especialmente desde que estas raíces pueden hallarse por procedimientos adecuados, con cualquier grado de aproximación. El arte de la resolución numérica de las ecuaciones es, naturalmente, muy importante y está muy desarrollado; pero el problema de Abel y Ruffini es completamente distinto: ¿puede encontrarse la solución utilizando *únicamente operaciones racionales y radicales?* Fué el deseo de alcanzar claridad completa sobre esta cuestión lo que inspiró el magnífico desarrollo del álgebra moderna y de la teoría de grupos, iniciado por Ruffini, Abel y Galois (1811-1832).

El problema de probar la imposibilidad de ciertas construcciones geométricas nos procura uno de los ejemplos más sencillos de esta

tendencia del álgebra. Mediante el uso de conceptos algebraicos podremos demostrar en este capítulo la imposibilidad de trisecar el ángulo, de construir el heptágono regular, o de duplicar el cubo, con la sola ayuda de la regla y el compás. (El problema de la cuadratura del círculo es mucho más difícil de tratar; véase pág. 152). Nuestro punto de partida no será la cuestión negativa de la imposibilidad de ciertas construcciones, sino, por el contrario, la cuestión positiva de cómo pueden caracterizarse completamente todos los problemas *construíbles*. Después de haber contestado a esta cuestión, será tarea fácil probar que los problemas mencionados caen fuera de esta categoría.

A los diecisiete años, Gauss investigó la constructibilidad de los «p-ágonos» regulares (polígonos de p lados), siendo p un número primo. Sólo se conocía entonces la construcción para $p = 3$ y $p = 5$; Gauss descubrió que el «p-ágono» regular era *construíble* si, y sólo si, p es un «número primo de Fermat»; es decir,

$$p = 2^{2^n} + 1.$$

Los primeros números de Fermat son 3, 5, 17, 257, 65 537 (véase página 33). Tan entusiasmado se sintió el joven Gauss por su descubrimiento, que renunció a su intención de hacerse filólogo y resolvió dedicar su vida a la matemática y sus aplicaciones. Siempre recordó la primera de sus grandes proezas con particular orgullo. Después de su muerte le fué erigida en Gotinga una estatua de bronce, y no pudo encontrarse honor más adecuado que el de dar a su pedestal la forma de un polígono regular de 17 lados.

Cuando se trata de construcciones geométricas no hay que olvidar nunca que el problema no es el de dibujar figuras en la práctica con cierto grado de exactitud, sino el de demostrar que sin otra ayuda que la regla y el compás la solución puede hallarse teóricamente, suponiendo que nuestros instrumentos tienen precisión ideal. Lo que Gauss demostró es que sus construcciones pueden realizarse en principio. Su teoría nada tiene que ver con el método más sencillo de realizarlas efectivamente o con los artificios que permitan simplificar y acortar el número de pasos necesarios; ésta es una cuestión de importancia teórica mucho menor. Desde un punto de vista práctico, ninguna construcción puede resultar tan satisfactoria como la que se obtiene mediante el uso de un buen trasportador de ángulos. El no entender adecuadamente el carácter teórico de la cuestión de las construcciones geométricas y la obstinación en querer desconocer los hechos científicos bien establecidos son responsables de la persistencia de los innumerables trisectores de ángulos y cuadradores de

círculos. Aquellos de entre éstos que sean capaces de entender las matemáticas elementales, pueden sacar provecho del estudio de este capítulo.

Una vez más recalcaremos que en algunas ocasiones nuestro concepto de construcción geométrica parece artificial. La regla y el compás son, ciertamente, los instrumentos más simples para dibujar; pero la restricción de utilizar exclusivamente estos instrumentos no es de forma alguna esencial en geometría. Como los matemáticos griegos sabían, tiempo ha, ciertos problemas —tales como el de duplicar el cubo— pueden resolverse si se permite, p. ej., utilizar una regla en forma de ángulo recto; es bastante fácil idear instrumentos distintos del compás, por medio de los cuales es posible dibujar elipses, hipérbolas y curvas mucho más complicadas, y cuyo uso amplía considerablemente el dominio de las figuras *construibles*. En los próximos párrafos, sin embargo, seguiremos adheridos al concepto usual de construcciones geométricas mediante el uso exclusivo de la regla y el compás.

PARTE PRIMERA

DEMOSTRACIONES DE IMPOSIBILIDAD Y ÁLGEBRA

I. CONSTRUCCIONES GEOMÉTRICAS FUNDAMENTALES

1. Construcción de cuerpos de números y extracción de raíces cuadradas.—Para dar forma a nuestras ideas generales comenzaremos examinando algunas de las construcciones clásicas. La clave de una comprensión más profunda reside en trasladar los problemas geométricos al lenguaje del álgebra.

Todo problema de construcción geométrica es del siguiente tipo: se da cierto conjunto de segmentos rectilíneos, a, b, c, ... y se pide construir uno o más segmentos, x, y, ... Es siempre posible formular los problemas de este modo, aunque a primera vista tengan un aspecto muy diferente. Los segmentos requeridos pueden aparecer como lados de un triángulo a construir, como radios de círculos, o como coordenadas rectangulares de ciertos puntos (véase, p. ej., pág. 138). Por sencillez, supondremos que sólo se pide un segmento x. La construcción geométrica equivale entonces a resolver un problema algebraico; primero debemos hallar una relación (ecuación) entre la cantidad pedida x y las dadas a, b, c, ...; después hay que hallar la cantidad buscada x resolviendo esta ecuación, y, finalmente, debemos determinar si esta solución puede obtenerse mediante un proceso algebraico que corresponda a las construcciones con regla y compás. Éste es el principio de la geometría analítica: la caracterización cuantitativa de los objetos geométricos mediante números reales, basada en la introducción del continuo numérico real, que es el fundamento de toda la teoría.

Observemos en primer lugar que algunas de las operaciones algebraicas más sencillas corresponden a construcciones geométricas elementales. Dados dos segmentos de longitudes a y b (medidos con un segmento «unidad» dado), es inmediato construir $a + b$, $a - b$, ra (donde r es cualquier número racional), a/b y ab.

Para construir $a + b$ (Fig. 27) trazamos una recta y llevamos con el compás las distancias $OA = a$ y $AB = b$; entonces $OB = a + b$. Análogamente, para $a - b$, llevamos $OA = a$ y $AB = b$; pero esta vez AB en sentido opuesto a OA; entonces $OB = a - b$. Para construir $3a$ sumamos simplemente $a + a + a$; de forma análoga, podemos construir pa, siendo p cualquier entero. Construiremos $a/3$ me-

diante el siguiente artificio (Fig. 28): llevamos $OA = a$ sobre una recta, y dibujamos una segunda recta por O. Sobre ésta llevamos un segmento arbitrario $OC = c$, y construímos $OD = 3c$. Unimos A con D, y trazamos desde C una recta paralela a AD, que corta a OA en B. Los triángulos OBC y OAD son semejantes; por tanto, $OB/a = OB/OA = OC/OD = 1/3$, y $OB = a/3$. Del mismo modo podemos construir a/q, donde q es cualquier entero. Aplicando esta operación al segmento pa podemos construir ra, siendo $r = p/q$ un número racional cualquiera.

Para construir a/b (Fig. 29) llevamos $OB = b$ y $OA = a$ sobre los lados de un ángulo O, y sobre OB llevamos $OD = 1$. Desde D trazamos una paralela a AB, que corta a OA en C. Entonces OC tendrá la longitud a/b. La construcción de ab se muestra en la figura 30, donde AD es una paralela a BC desde A. De estas consideraciones resulta que los *procesos algebraicos «racionales»*—adición, sustracción, multiplicación y división de cantidades conocidas—*pueden efectuarse por medio de construcciones geométricas*. A partir de segmentos dados, medidos por números reales a, b, c, ... podemos, por sucesivas aplicaciones de estas sencillas construcciones, construir cualquier cantidad que sea expresable mediante a, b, c, ... en forma racional; es decir, por medio de la aplicación reiterada de adiciones, restas, multiplicaciones y divisiones. La totalidad de las cantidades que pueden obtenerse en esta forma a partir de a, b, c, ... constituye lo que se llama

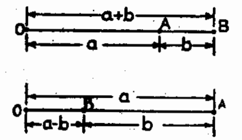


FIG. 27.—Construcción de $a+b$ y de $a-b$.

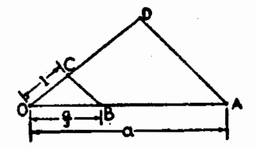


FIG. 28.—Construcción de $a/3$.

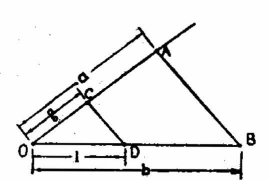


FIG. 29.—Construcción de a/b.

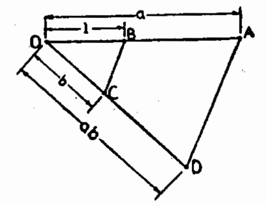


FIG. 30.—Construcción de ab.

un *cuerpo de números*, un conjunto de números tal que toda operación racional aplicada a dos o más miembros del conjunto da a su vez lugar a un número del conjunto. Recordemos que los números racionales, los números reales y los números complejos son ejemplos de cuerpos de números. En el caso presente, el cuerpo se dice *engendrado* por los números dados a, b, c, ...

La nueva construcción decisiva, que nos lleva fuera del cuerpo así obtenido, es la extracción de una raíz cuadrada; dado un segmento a, $\sqrt{a}$ puede también construirse utilizando sólo la regla y el compás. Sobre una recta llevamos $OA = a$ y $AB = 1$ (Fig. 31). Trazamos una circunferencia con el segmento OB como diámetro y después la perpendicular a OB desde A, la cual corta a la circunferencia en C. El triángulo OBC tiene un ángulo recto en C, ya que es sabido por geometría elemental que el ángulo inscrito en una semicircunferencia es recto.

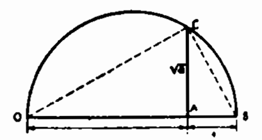


FIG. 31.—Construcción de $\sqrt{a}$.

Luego $\widehat{OCA} = \widehat{ABC}$ por ser semejantes los triángulos rectángulos OAC y CAB, y tenemos, para $x = AC$,

$$\frac{a}{x} = \frac{x}{1}, \quad x^2 = a, \quad x = \sqrt{a}.$$

2. Polígonos regulares.—Vamos a considerar ahora algunos problemas de construcción algo más complicados. Comencemos por el *decágono regular*. Supongamos que un decágono regular está inscrito en un círculo de radio 1 (Fig. 32) y llamemos x a su lado. Dado que x subtiende un ángulo central de 36º, los otros dos ángulos del triángulo deben valer cada uno 72º, y, por tanto, la recta de puntos que biseca al ángulo A divide al triángulo OAB en dos triángulos isósceles, cada uno con dos lados iguales de longitud x. El radio del círculo se ha dividido así en dos segmentos, x y $1 - x$. Por ser OAB semejante al triángulo isósceles menor se tiene $1/x = x/(1 - x)$. De esta proporción deducimos la ecuación cuadrática $x^2 + x - 1 = 0$, una de cuyas soluciones es $x = (\sqrt{5} - 1)/2$. (La otra solución debe desecharse por ser negativa.) De esto resulta evidente que se puede construir x geométricamente. Teniendo la longitud x podemos ahora construir el decágono regular, llevando su longitud como cuerda del círculo. El pentágono regular puede obtenerse sin más que unir de dos en dos los vértices del decágono regular.

Además de la construcción de $\sqrt{5}$ por el método de la figura 31, podemos

también obtenerlo como hipotenusa del triángulo rectángulo de catetos 1 y 2. Resulta x al restar la unidad de $\sqrt{5}$ y dividir el resultado por 2.

Los matemáticos griegos llamaban a la razón $OB : AB$ del problema precedente razón áurea, pues consideraban que un rectángulo cuyos lados estuviesen en esta relación era el más agradable estéticamente. Su valor, dicho sea de paso, es 1,62, aproximadamente.

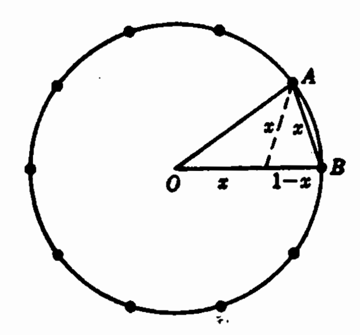


FIG. 32.—Decágono regular.

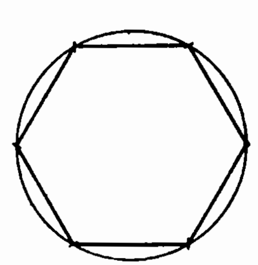


FIG. 33.—Hexágono regular.

De todos los polígonos regulares, el hexágono es el más sencillo de construir. Tracemos un círculo de radio r; la longitud del lado del hexágono regular inscrito en este círculo será entonces igual a r. El hexágono puede construirse llevando a partir de un punto de la circunferencia cuerdas de longitud r hasta obtener los seis vértices.

Del n-ágono regular podemos obtener el $2n$-ágono regular bisecando el arco subtendido en la circunferencia circunscrita por cada lado del n-ágono, utilizando también los puntos adicionales así obtenidos como vértices originales del $2n$-ágono pedido. Partiendo del diámetro de la circunferencia (ó «2-ágono»), podemos, por tanto, construir los polígonos de 4, 8, 16, ..., 2^n lados. Análogamente, es posible obtener los de 12, 24, 48 lados, etc., a partir del hexágono, y los de 20 y 40 lados, etc., partiendo del decágono.

Si s_n designa la longitud del lado del n-ágono regular inscrito en el círculo unidad (círculo de radio 1), entonces el lado del $2n$-ágono regular tiene la longitud

$$s_{2n} = \sqrt{2 - \sqrt{4 - s_n^2}}.$$

Esto puede demostrarse como sigue: en la figura 34, s_n es igual a $DE = 2DC$; $s_{2n} = DB$, y $AB = 2$. El área del triángulo rectángulo ABD es $^1/_2 BD \cdot AD =$

$= \frac{1}{2}AB \cdot CD$. Como $AD = \sqrt{AB^2 - DB^2}$, sustituyendo $AB = 2$, $BD = s_{2n}$, $CD = \frac{1}{2}s_n$, e igualando las dos expresiones del área, resulta

$$s_n = s_{2n}\sqrt{4 - s_{2n}^2} \qquad \text{o} \qquad s_n^2 = s_{2n}^2(4 - s_{2n}^2).$$

Resolviendo esta ecuación cuadrática en $x = s_{2n}^2$ y observando que x debe ser menor que 2, se llega fácilmente a la fórmula anterior.

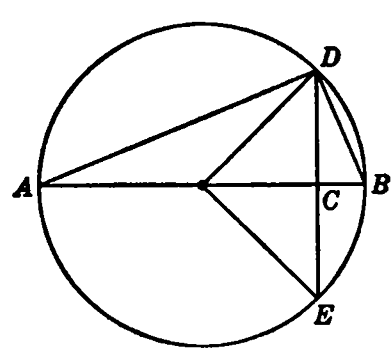


FIG. 34.

De esta fórmula y del hecho de que s_4 (lado del cuadrado) sea igual a $\sqrt{2}$ se deduce:

$$s_8 = \sqrt{2 - \sqrt{2}} \qquad s_{16} = \sqrt{2 - \sqrt{2 + \sqrt{2}}},$$

$$s_{32} = \sqrt{2 - \sqrt{2 + \sqrt{2 + \sqrt{2}}}}, \text{ etc.}$$

Como fórmula general obtenemos para $n > 2$:

$$s_{2^n} = \sqrt{2 - \sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}$$

que incluye $n - 1$ raíces cuadradas sucesivas. El perímetro del 2^n-ágono regular inscrito será $2^n s_{2^n}$. Al tender n a infinito, el 2^n-ágono tiende a confundirse con la circunferencia, y, en consecuencia, $2^n s_{2^n}$ tiende a la longitud de la circunferencia del círculo unidad, que por definición es 2π. Obtenemos así, sustituyendo m por $n - 1$ y suprimiendo el factor 2, la fórmula asintótica para π:

$$2^m \underbrace{\sqrt{2 - \sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}}_{m \text{ raíces cuadradas}} \to \pi \text{ cuando } m \to \infty.$$

Ejercicio: Puesto que $2^m \to \infty$ demuéstrese que

$$\underbrace{\sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}_{n \text{ raíces cuadradas}} \to 2 \text{ cuando } n \to \infty.$$

Los resultados obtenidos poseen el siguiente rasgo característico: *Los lados del 2^n-ágono, del $5 \cdot 2^n$-ágono, y del $3 \cdot 2^n$-ágono, pueden construirse mediante procesos de sumas, restas, productos, divisiones y extracciones de raíces cuadradas.*

***3. Problema de Apolonio.**—Otro problema de construcción, muy sencillo desde el punto de vista algebraico, es el famoso problema ya mencionado de los círculos tangentes de Apolonio. En lo que sigue no nos es necesario encontrar una construcción particularmente elegante; lo que importa es que, en principio, el problema pueda resolverse con regla y compás solamente. Daremos una breve indicación de la demostración y más adelante (véase pág. 173) veremos un método constructivo más elegante.

Supongamos que los centros de las circunferencias dadas tengan coordenadas (x_1, y_1), (x_2, y_2) y (x_3, y_3), y radios r_1, r_2 y r_3, respectivamente. Designemos el centro y el radio del círculo pedido por (x, y) y r. Entonces, la condición para que dicho círculo sea tangente a los tres dados se obtiene observando que la distancia entre los centros de dos círculos tangentes es igual a la suma o diferencia de los radios, según que éstos sean tangentes exterior o interiormente. Esto nos da las ecuaciones:

$$(x - x_1)^2 + (y - y_1)^2 - (r \pm r_1)^2 = 0, \qquad [1]$$
$$(x - x_2)^2 + (y - y_2)^2 - (r \pm r_2)^2 = 0, \qquad [2]$$
$$(x - x_3)^2 + (y - y_3)^2 - (r \pm r_3)^2 = 0, \qquad [3]$$

o bien,

$$x^2 + y^2 - r^2 - 2xx_1 - 2yy_1 \pm 2rr_1 + x_1^2 + y_1^2 - r_1^2 = 0, \qquad [1a]$$

etcétera. El signo $\pm$ debe elegirse en cada una de estas ecuaciones según que las circunferencias sean tangentes exterior o interiormente (véase Fig. 35). Las ecuaciones [1], [2], [3], son cuadráticas con tres incógnitas, x, y, r, y en las tres los términos de segundo grado son los mismos, como se ve en la forma desarrollada [1 a]. Por tanto, restando [2] de [1], obtendremos una ecuación lineal en x, y, r:

$$ax + by + cr = d, \qquad [4]$$

donde $a = 2(x_2 - x_1)$, etc. Análogamente, restando [3] de [1] tenemos otra ecuación lineal,

$$a'x + b'y + c'r = d'. \qquad [5]$$

Resolviendo [4] y [5] respecto a x e y en función de r, y sustituyendo en [1], tendremos una ecuación cuadrática en r, que puede

resolverse por operaciones racionales y extracción de una raíz cuadrada (véase pág. 101). Habrá, en general, dos soluciones de esta ecuación, de las cuales sólo una es positiva. Después de determinar r mediante esta ecuación, obtendremos x e y de las dos ecuaciones lineales [4] y [5]. El círculo con centro (x, y) y radio r será tangente a los tres círculos dados. En todo el proceso hemos usado sólo operaciones ra-

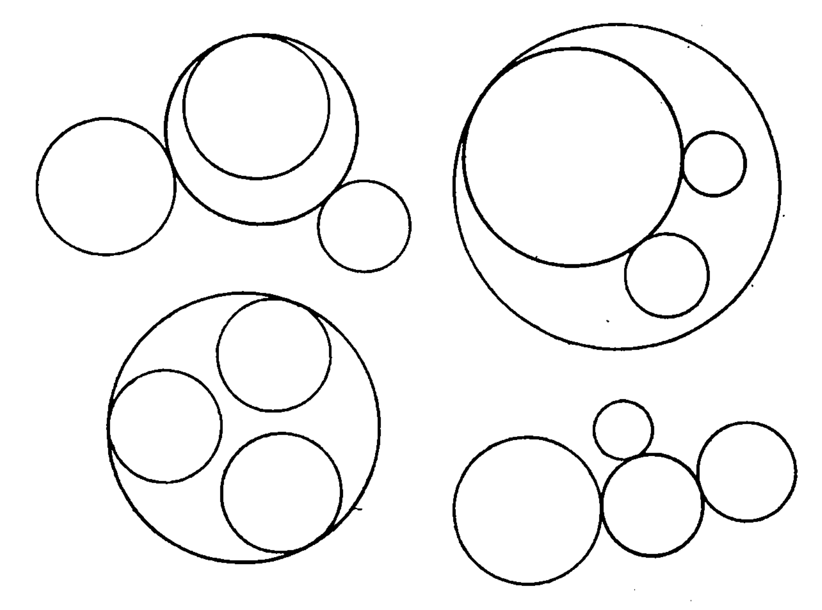


FIG. 35.—Círculos de Apolonio.

cionales y extracciones de raíces cuadradas. Síguese de ello que r, x e y pueden ser construídos con regla y compás.

Habrá, en general, ocho soluciones del problema de Apolonio, correspondientes a las $2 \cdot 2 \cdot 2 = 8$ combinaciones posibles de los signos más y menos en las ecuaciones [1], [2] y [3]. Estas elecciones corresponden a las condiciones de que las circunferencias pedidas sean tangentes exterior o interiormente a cada uno de los tres círculos dados. Puede ocurrir que nuestro proceso algebraico no nos dé valores reales para x, y y r. Esto sucederá, p. ej., si los tres círculos dados son concéntricos, no existiendo entonces solución geométrica del problema. También cabe esperar posibles *degeneraciones* de la solución, como en el caso en que los tres círculos dados degeneren en tres puntos de una recta. Entonces, el círculo de Apolonio degenera en esta recta. No

discutiremos estas posibilidades con detalle; el lector con alguna experiencia de álgebra podrá completar el análisis.

*II. NÚMEROS CONSTRUÍBLES Y CUERPOS DE NÚMEROS

1. Teoría general.—En nuestra discusión previa queda indicado el fondo algebraico general de las construcciones geométricas. Cada construcción con regla y compás consiste en una sucesión de operaciones de las enumeradas a continuación: 1) unir dos puntos por una recta; 2) hallar el punto de intersección de dos rectas; 3) trazar una circunferencia de radio y centro dados; 4) hallar los puntos de intersección de una circunferencia con otra circunferencia o con una recta. Un elemento (punto, recta, circunferencia) se considera conocido si se da desde el principio o si ha sido construído en algún paso previo. Para un análisis teórico, podemos referir la construcción en conjunto a un sistema de coordenadas x, y (véase pág. 82). Los elementos dados pueden representarse por puntos o segmentos en el plano x, y. Si sólo se da un segmento inicial, podemos tomarlo como unidad de longitud, lo que determina el punto $x = 1$, $y = 0$. A veces, aparecen elementos «arbitrarios»: se trazan líneas arbitrarias, se eligen puntos o radios arbitrarios. (Un ejemplo de elemento arbitrario aparece al construir el punto medio de un segmento: dibujamos dos circunferencias de radios iguales, pero arbitrarios, con sus centros en los extremos del segmento, y unimos sus intersecciones.) En tales casos, elegimos el elemento de manera que sea racional; es decir, los puntos arbitrarios, con coordenadas x, y racionales; las rectas arbitrarias $ax+by+c=0$, con coeficientes a, b, c racionales; los círculos arbitrarios, con centros de coordenadas racionales y radios racionales. Haremos siempre la elección de los elementos arbitrarios de manera que sean racionales; si los elementos son verdaderamente arbitrarios, esta restricción no puede afectar al resultado de la construcción.

Por sencillez, supondremos en la siguiente discusión que sólo se da un elemento inicial, el segmento de longitud 1. Según § I podemos construir con regla y compás todos los números que puedan deducirse de la unidad mediante procesos racionales de suma, resta, multiplicación y división; es decir, todos los números racionales r/s, siendo r y s enteros. El sistema de los números racionales es «cerrado» respecto a las operaciones racionales; esto es, la suma, diferencia, producto o cociente de dos números racionales—excluyendo, como siempre, la división por cero—es también un número racional. Todo con-

junto de números que posea esta propiedad de ser cerrado respecto a las cuatro operaciones, se denomina *cuerpo de números*.

Ejercicio: Demuéstrese que todo cuerpo contiene al menos todos los números racionales. (*Indicación:* Si $a \neq 0$ es un número del cuerpo F, entonces $a/a = 1$ pertenece a F, y a partir de 1 se obtiene todo número racional por operaciones racionales.)

Partiendo de la unidad, podemos construir el cuerpo completo de los números racionales y, en consecuencia, todos los puntos racionales (es decir, los puntos con ambas coordenadas racionales) del plano x, y. Podemos obtener nuevos números, irracionales, haciendo uso del compás para construir, p. ej., el número $\sqrt{2}$, que, como sabemos por el capítulo II, § II, no pertenece al cuerpo racional. Habiendo construído $\sqrt{2}$ mediante las construcciones «racionales» de § I, podemos hallar todos los números de la forma

$$a + b\sqrt{2}, \qquad [1]$$

siendo a y b racionales y, por tanto, construíbles. También podemos construir todos los números de la forma

$$\frac{a + b\sqrt{2}}{c + d\sqrt{2}} \qquad \text{o} \qquad (a + b\sqrt{2})(c + d\sqrt{2}),$$

donde a, b, c y d son racionales. Por otra parte, estos números pueden escribirse en la forma [1]; en efecto:

$$\frac{a + b\sqrt{2}}{c + d\sqrt{2}} = \frac{a + b\sqrt{2}}{c + d\sqrt{2}} \cdot \frac{c - d\sqrt{2}}{c - d\sqrt{2}} =$$

$$= \frac{ac - 2bd}{c^2 - 2d^2} + \frac{bc - ad}{c^2 - 2d^2}\sqrt{2} = p + q\sqrt{2},$$

siendo p y q racionales. (El denominador $c^2 - 2d^2$ no puede ser cero, pues si $c^2 - 2d^2 = 0$, entonces $\sqrt{2} = c/d$, lo que contradice la demostrada irracionalidad de $\sqrt{2}$.) Igualmente

$$(a + b\sqrt{2})(c + d\sqrt{2}) = (ac + 2bd) + (bc + ad)\sqrt{2} = r + s\sqrt{2},$$

siendo r y s racionales. De aquí que todo cuanto podemos obtener con la construcción de $\sqrt{2}$ es el conjunto de números de la forma [1], siendo a y b racionales y arbitrarios.

Ejercicio: Siendo $p = 1 + \sqrt{2}$, $q = 2 - \sqrt{2}$, $r = -3 + \sqrt{2}$ pónganse bajo la forma [1] los números

$$\frac{p}{q},\ p + p^2,\ (p - p^2)\frac{q}{r},\ \frac{pqr}{1 + r^2},\ \frac{p + qr}{q + pr^2}$$

Estos números [1] forman un *cuerpo*, como muestra la discusión precedente. (Es obvio que la suma y diferencia de dos números de la forma [1] tiene también la misma forma.) Este cuerpo es más amplio que el cuerpo racional, que es una parte o *subcuerpo* de él. Pero, naturalmente, es menos amplio que el cuerpo de *todos* los números reales. Llamemos al cuerpo racional F_0, y al nuevo cuerpo de números de la forma [1], F_1. La posibilidad de construir cualquier número del «cuerpo ampliado» F_1 ha sido ya establecida. Podemos ahora extender el alcance de nuestras construcciones, p. ej., tomando un número de F_1, tal como $k = 1 + \sqrt{2}$, y extrayendo la raíz cuadrada; obtenemos así el número construíble

$$\sqrt{1 + \sqrt{2}} = \sqrt{k},$$

y con él, de acuerdo con § I, el cuerpo formado por todos los números

$$p + q\sqrt{k}, \qquad [2]$$

donde ahora p y q son números arbitrarios de F_1; es decir, de la forma $a + b\sqrt{2}$, con a, b de F_0, esto es, racionales.

Ejercicio: Represéntense en la forma [2] los números

$$(\sqrt{k})^3,\ \frac{1 + (\sqrt{k})^3}{1 + \sqrt{k}},\ \frac{\sqrt{2}\sqrt{k} + \frac{1}{\sqrt{2}}}{(\sqrt{k})^3 - 3},\ \frac{(1 + \sqrt{k})(2 - \sqrt{k})\left(\sqrt{2} + \frac{1}{\sqrt{k}}\right)}{1 + \sqrt{2}k}$$

Todos estos números han sido construídos en la hipótesis de un solo segmento dado inicialmente. Si se dan dos, elegiremos uno de ellos como unidad de longitud, y supondremos que la longitud del otro segmento medido con esta unidad es α. Entonces podemos construir el cuerpo G formado por todos los números de la forma

$$\frac{a_m\alpha^m + a_{m-1}\alpha^{m-1} + \cdots + a_1\alpha + a_0}{b_n\alpha^n + b_{n-1}\alpha^{n-1} + \cdots + b_1\alpha + b_0}$$

donde los números $a_0, \cdots, a_m$ y $b_0, \cdots, b_n$ son racionales, y m y n enteros positivos.

Ejercicio: Dados los segmentos de longitudes 1 y α, constrúyase $1 + \alpha + \alpha^2$, $(1 + \alpha)/(1 - \alpha)$, α^3.

Supongamos ahora, con mayor generalidad, que somos capaces de construir todos los números de un cuerpo F. Vamos a demostrar que *el uso de la regla sola no nos permitirá salir fuera del cuerpo* F. La ecuación de la recta que une los puntos de coordenadas a_1, b_1 y a_2, b_2, pertenecientes a F, es $(b_1 - b_2)x + (a_2 - a_1)y + (a_1b_2 - a_2b_1) = 0$ (véase pág. 501); sus coeficientes son expresiones racionales formadas por números de F, y, por definición de cuerpo, también pertenecen a F. Además, si tenemos dos rectas $\alpha x+\beta y-\gamma=0$ y $\alpha' x+\beta' y-\gamma'=0$, con coeficientes de F, las coordenadas de su punto de intersección, halladas resolviendo este sistema de dos ecuaciones, son:

$$x = \frac{\gamma\beta' - \beta\gamma'}{\alpha\beta' - \beta\alpha'}, \quad y = \frac{\alpha\gamma' - \gamma\alpha'}{\alpha\beta' - \beta\alpha'}$$

Como éstos son también números de F, es evidente que el uso de la regla sola no nos permitirá salir fuera de los confines del cuerpo F.

Ejercicios: Las rectas $x + \sqrt{2}y - 1 = 0$, $2x - y + \sqrt{2} = 0$, tienen coeficientes del cuerpo [1]. Calcúlense las coordenadas de su punto de intersección, y verifíquese que tienen la forma [1].

Únanse los puntos $(1, \sqrt{2})$ y $(\sqrt{2}, 1 - \sqrt{2})$ mediante la recta $ax + by + c = 0$ y verifíquese que los coeficientes son de la forma [1].

Hágase lo mismo respecto al cuerpo [2] para las rectas

$$\sqrt{1+\sqrt{2}}\,x + \sqrt{2}\,y = 1, \quad (1+\sqrt{2})\,x - y = 1 - \sqrt{1+\sqrt{2}},$$

y los puntos $(\sqrt{2}, -1)$, $(1+\sqrt{2}, \sqrt{1+\sqrt{2}})$, respectivamente.

Sólo podemos salir de los confines de F haciendo uso del compás. Para conseguir esto, elegimos un elemento k de F tal que $\sqrt{k}$ no pertenezca a F. Entonces podemos construir $\sqrt{k}$ y, por tanto, todos los números

$$a + b\sqrt{k}, \qquad [3]$$

donde a y b son racionales, o incluso elementos arbitrarios de F. La suma y diferencia de dos números $a + b\sqrt{k}$ y $c + d\sqrt{k}$, su producto, $(a + b\sqrt{k})(c + d\sqrt{k}) = (ac + kbd) + (ad + bc)\sqrt{k}$, y su cociente,

$$\frac{a + b\sqrt{k}}{c + d\sqrt{k}} = \frac{(a + b\sqrt{k})(c - d\sqrt{k})}{c^2 - kd^2} = \frac{ac - kbd}{c^2 - kd^2} + \frac{bc - ad}{c^2 - kd^2}\sqrt{k},$$

son de nuevo de la forma $p + q\sqrt{k}$, siendo p y q de F. (El denominador $c^2 - kd^2$ no puede anularse, salvo que c y d sean ambos cero;

en caso contrario $\sqrt{k} = c/d$, sería un número de F, mientras hemos supuesto que $\sqrt{k}$ no pertenecía a F.) Luego el conjunto de números de la forma $a + b\sqrt{k}$ forma un cuerpo F', que contiene el cuerpo original F, pues podemos, en particular, elegir $b = 0$. F' se llama *cuerpo generalizado* o *extensión* de F, y F *subcuerpo* de F'.

Como ejemplo, sea F el cuerpo $a + b\sqrt{2}$, con a y b racionales y $k = \sqrt{2}$. Entonces, los números del cuerpo generalizado F' se representan por $p + q\sqrt[4]{2}$, siendo p y q de F, $p = a + b\sqrt{2}$, $q = a' + b'\sqrt{2}$, con a, b, a' y b' racionales. Todo número de F' puede ser reducido a esta forma; p. ej.,

$$\frac{1}{\sqrt{2} + \sqrt[4]{2}} = \frac{\sqrt{2} - \sqrt[4]{2}}{(\sqrt{2} + \sqrt[4]{2})(\sqrt{2} - \sqrt[4]{2})} = \frac{\sqrt{2} - \sqrt[4]{2}}{2 - \sqrt{2}} =$$

$$= \frac{\sqrt{2}}{2 - \sqrt{2}} - \frac{\sqrt[4]{2}}{2 - \sqrt{2}} = \frac{\sqrt{2}\,(2 + \sqrt{2})}{4 - 2} - \frac{(2 + \sqrt{2})}{4 - 2}\sqrt[4]{2} =$$

$$= (1 + \sqrt{2}) - (1 + {}^1/_2\sqrt{2})\sqrt[4]{2}.$$

Ejercicio: Sea F el cuerpo $p + q\sqrt{2 + \sqrt{2}}$, donde p y q son de la forma $a + b\sqrt{2}$, y a y b racionales. Represéntese en esa forma el número

$$\frac{1 + \sqrt{2 + \sqrt{2}}}{2 - 3\sqrt{2 + \sqrt{2}}}$$

Hemos visto que si partimos de un cuerpo F de números *construibles* que contiene el número k, entonces, por medio de la regla y una única utilización del compás podemos construir $\sqrt{k}$ y, en consecuencia, todos los números de la forma $a + b\sqrt{k}$, en tanto que a y b sean de F.

Recíprocamente, probaremos ahora que, mediante una sola aplicación del compás, podemos obtener *sólo* números de esta forma. Pues lo que el compás realiza en una construcción es definir puntos (o sus coordenadas) como intersecciones de una circunferencia y una recta, o de dos circunferencias. La circunferencia de centro ξ, η y radio r tiene como ecuación $(x - \xi)^2 + (y - \eta)^2 = r^2$; luego si ξ, η y r pertenecen a F, la ecuación podrá escribirse en la forma:

$$x^2 + y^2 + 2\alpha x + 2\beta y + \gamma = 0,$$

con coeficientes α, β y γ de F. Una recta $ax + by + c = 0$, que une dos puntos cuyas coordenadas están en F, tiene coeficientes a, b, c de F, como ya hemos visto (pág. 141). Eliminando y entre estas dos ecuaciones obtenemos para la abscisa x de un punto de intersección de la circunferencia y la recta una ecuación cuadrática de la forma $Ax^2 + Bx + C = 0$, cuyos coeficientes A, B, C, de F, son $A = a^2 + b^2$, $B = 2(ac + b^2\alpha - ab\beta)$, $C = c^2 - 2bc\beta + b^2\gamma$. La solución viene dada por la fórmula

$$x = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A},$$

que es de la forma $p + q\sqrt{k}$, con p, q, k de F. Una fórmula análoga nos da la ordenada y del punto de intersección.

Por otra parte, si tenemos dos circunferencias:

$$\begin{aligned} x^2 + y^2 + 2\alpha x + 2\beta y + \gamma &= 0, \\ x^2 + y^2 + 2\alpha' x + 2\beta' y + \gamma' &= 0, \end{aligned}$$

restando la segunda ecuación de la primera, obtenemos la ecuación lineal

$$2(\alpha - \alpha')x + 2(\beta - \beta')y + (\gamma - \gamma') = 0,$$

que puede resolverse junto con la ecuación de la primera circunferencia, como antes. En uno y otro caso, la construcción da las coordenadas x e y del punto o de los dos nuevos puntos, y estas cantidades son de la forma $p + q\sqrt{k}$, con p, q, k de F. En particular, naturalmente, $\sqrt{k}$ puede pertenecer a F; p. ej., cuando sea $k = 4$. Entonces la construcción no proporciona nada esencialmente nuevo, y no salimos de F. Pero éste no es el caso general.

Ejercicio: Considérese el círculo de radio $2\sqrt{2}$ con centro en el origen, y la recta que une los puntos $(1/2, 0)$, $(4\sqrt{2}, \sqrt{2})$. Hállese el cuerpo F' determinado por las coordenadas de los puntos de intersección de la circunferencia y la recta. Hágase lo mismo respecto de la intersección de la circunferencia dada con la de radio $\sqrt{2}/2$ y centro $(0, 2\sqrt{2})$.

Resumiendo otra vez: dadas ciertas cantidades iniciales, podemos construir sólo con la regla todas las cantidades del cuerpo F engendrado mediante procesos racionales a partir de las cantidades dadas. Utilizando el compás, es posible extender el cuerpo F de las cantidades construíbles a un cuerpo más amplio, eligiendo un número k de F, extrayendo su raíz cuadrada y construyendo el cuerpo F' formado por todos los números de la forma $a + b\sqrt{k}$, donde a y b son de F.

Éste es un subcuerpo de F', porque todas las cantidades de F están también contenidas en F', ya que en la expresión $a + b\sqrt{k}$ podemos elegir $b = 0$. (Se supone que $\sqrt{k}$ es un nuevo número no perteneciente a F, pues de otro modo el proceso de adjunción de $\sqrt{k}$ no haría variar la situación, y F' sería idéntico a F.) Hemos probado que todo paso en una construcción geométrica (trazado de la recta que une dos puntos, de la circunferencia de centro y radio dados, o determinación de la intersección de dos rectas o circunferencia conocidas) puede, bien producir cantidades del cuerpo ya conocido o, mediante la construcción de una raíz cuadrada, dar lugar a un nuevo cuerpo ampliado de números *construíbles*.

La totalidad de los números *construíbles* puede ser descrita ahora con precisión. Partimos de un cuerpo dado F_0, definido por cantidades iniciales dadas; p. ej., el cuerpo de todos los números racionales, si sólo se da un segmento, elegido como unidad. A continuación, mediante la adjunción de $\sqrt{k_0}$, donde k_0 es de F_0, pero no $\sqrt{k_0}$, formamos el cuerpo ampliado F_1 de números *construíbles*, que consta de todos los números de la forma $a_0 + b_0\sqrt{k_0}$, en que a_0 y b_0 son números cualesquiera de F_0. Después se define F_2, nueva extensión de F_1, como conjunto de todos los números de la forma $a_1 + b_1\sqrt{k_1}$, siendo a_1 y b_1 números de F_1, y k_1 un número de F_1 cuya raíz cuadrada no está en F_1. Repitiendo este proceso, podemos obtener un cuerpo F_n después de n adjunciones de raíces cuadradas. *Números construíbles son aquellos y sólo aquellos que pueden hallarse mediante una tal sucesión de cuerpos ampliados, esto es, que pertenecen a un cuerpo* F_n *del tipo descrito.* La magnitud del número n de extensiones necesarias no importa; sólo mediría el grado de complejidad del problema.

El siguiente ejemplo puede aclarar el proceso. Deseamos obtener el número

$$\sqrt{6 + \sqrt{\sqrt{\sqrt{1 + \sqrt{2}} + \sqrt{3}} + 5}}.$$

Sea F_0 el cuerpo de los números racionales. Hagamos $k_0 = 2$, obteniendo el cuerpo F_1, que contiene el número $1 + \sqrt{2}$. Ahora tomamos $k_1 = 1 + \sqrt{2}$ y $k_2 = 3$. Por supuesto, 3 está en el cuerpo original F_0, y *a fortiori* en el cuerpo F_2, por lo cual es perfectamente lícito elegir $k_2 = 3$. Luego tomamos $k_3 = \sqrt{1 + \sqrt{2}} + \sqrt{3}$, y, finalmente, $k_4 = \sqrt{\sqrt{1 + \sqrt{2}} + \sqrt{3}} + 5$. El cuerpo F_5 así construído contiene el número deseado, pues $\sqrt{6}$ está a su vez en F_5, ya que $\sqrt{2}$ y $\sqrt{3}$, y en consecuencia su producto, están en F_3 y también, por tanto, en F_5.

Ejercicios: Verifíquese que, partiendo del cuerpo racional, el lado del 2^m-ágono regular (véase pág. 135) es un número *construíble*, con $n = m - 1$. Determínese la sucesión de cuerpos generalizados. Hágase lo mismo con los números

$$\sqrt{1+\sqrt{2}+\sqrt{3}+\sqrt{5}}, \qquad (\sqrt{5}+\sqrt{11})/(1+\sqrt{7-\sqrt{3}}),$$
$$(\sqrt{2+\sqrt{3}})\,(\sqrt[3]{2}+\sqrt{1+\sqrt{2+\sqrt{5}}+\sqrt{3-\sqrt{7}}}).$$

3. Todos los números construíbles son algebraicos.—Si el cuerpo F_0, inicial, es el cuerpo racional engendrado por un único segmento, entonces todos los números construíbles son algebraicos (véase pág. 112). Los números del cuerpo F_1 son raíces de ecuaciones cuadráticas, los de F_2, raíces de ecuaciones de cuarto grado, y, en general, los números de F_k son raíces de ecuaciones de grado 2^k, con coeficientes racionales. Para probar esto para el cuerpo F_2, consideremos como ejemplo $x = \sqrt{2} + \sqrt{3+\sqrt{2}}$. Tenemos $(x - \sqrt{2})^2 = 3 + \sqrt{2}$; $x^2 + 2 - 2\sqrt{2}x = 3 + \sqrt{2}$, ó $x^2 - 1 = \sqrt{2}\,(2x + 1)$, ecuación cuadrática con coeficientes del cuerpo F_1. Si elevamos al cuadrado, obtenemos finalmente

$$(x^2 - 1)^2 = 2(2x + 1)^2,$$

que es una ecuación de cuarto grado con coeficientes racionales.

En general, todo número del cuerpo F_2 tendrá la forma

$$x = p + q\sqrt{w}, \qquad [4]$$

donde p, q y w pertenecen al cuerpo F_1 y, por ello, son de la forma $p = a + b\sqrt{s}$, $q = c + d\sqrt{s}$, $w = e + f\sqrt{s}$, siendo a, b, c, d, e, f y s racionales. De [4] tenemos

$$x^2 - 2px + p^2 = q^2w,$$

cuyos coeficientes pertenecen todos al cuerpo F_1, engendrado por $\sqrt{s}$. Por tanto, esta ecuación puede escribirse en la forma

$$x^2 + ux + v = \sqrt{s}(rx + t),$$

siendo r, s, t, u y v racionales. Elevando al cuadrado ambos miembros, obtenemos una ecuación de cuarto grado

$$(x^2 + ux + v)^2 = s(rx + t)^2 \qquad [5]$$

con coeficientes racionales, según queríamos ver.

Ejercicios: 1. Hállense ecuaciones con coeficientes racionales para:
a) $x = \sqrt{2+\sqrt{3}}$; b) $x = \sqrt{2} + \sqrt{3}$; c) $x = 1/\sqrt{5+\sqrt{3}}$.

2. Hállense por métodos análogos ecuaciones de octavo grado para:
a) $x = \sqrt{2+\sqrt{2+\sqrt{2}}}$; b) $x = \sqrt{2} + \sqrt{1+\sqrt{3}}$; c) $x = 1 + \sqrt{5+\sqrt{3+\sqrt{2}}}$.

Para demostrar el teorema general si x es un cuerpo F_k con k arbitrario, probaríamos, mediante el procedimiento usado antes, que x satisface a una ecuación cuadrática con coeficientes de F_{k-1}. Repitiendo este procedimiento encontraríamos que x satisface a una ecuación de grado $2^2 = 4$ con coeficientes de F_{k-2}, etc.

Ejercicio: Complétese por inducción la demostración general para probar que x satisface a una ecuación de grado 2^l con coeficientes de F_{k-l}, siendo, $0 < l < k$. Este enunciado para $l = k$ es el teorema deseado.

*III. IRRESOLUBILIDAD DE LOS TRES PROBLEMAS GRIEGOS

1. Duplicación del cubo.—Estamos ahora en condiciones de investigar los viejos problemas de trisección del ángulo, duplicación del cubo y construcción del heptágono regular. Comenzaremos por el primero; si se da un cubo cuya arista es la unidad de longitud, su volumen será la unidad de volumen; se desea encontrar la arista x de un cubo cuyo volumen sea el doble. Dicha arista satisfará a la sencilla ecuación cúbica:

$$x^3 - 2 = 0. \qquad [1]$$

Nuestra demostración de que el número x no puede construirse con regla y compás es indirecta. Supondremos de momento que tal construcción es posible. Según lo dicho antes, esto significa que x pertenece a algún cuerpo F_k deducido, como ya se vió, del cuerpo racional, mediante sucesivas ampliaciones obtenidas por adjunción de raíces cuadradas. Como veremos, esta hipótesis nos lleva a una consecuencia absurda.

Sabemos ya que x no pertenece al cuerpo racional F_0, pues $\sqrt[3]{2}$ es un número irracional (Ej. 1, pág. 69). Luego x puede pertenecer sólo a algún cuerpo F_k, siendo k un entero positivo. Podemos suponer que k es el *menor* número entero positivo tal que x esté en algún F_k. Resulta entonces que x puede escribirse en la forma

$$x = p + q\sqrt{w},$$

donde p, q y w pertenecen a algún F_{k-1}, pero $\sqrt{w}$ no. A continuación, mediante un simple pero importante tipo de razonamiento algebraico, probaremos que si $x = p + q\sqrt{w}$ es una solución de la ecuación cúbica [1], entonces $y = p - q\sqrt{w}$ es también solución. Como x está en el cuerpo F_k, también x^3 y $x^3 - 2$ estarán en F_k, y tendremos

$$x^3 - 2 = a + b\sqrt{w}. \qquad [2]$$

en donde a y b pertenecen a F_{k-1}. Mediante un sencillo cálculo resulta $a = p^3 + 3pq^2w - 2$, $b = 3p^2q + q^3w$. Si hacemos

$$y = p - q\sqrt{w},$$

y sustituímos q por $-q$ en estas expresiones de a y b, vemos que

$$y^3 - 2 = a - b\sqrt{w}. \qquad [2']$$

Supongamos ahora que x es raíz de $x^3 - 2 = 0$, y, por tanto,

$$a + b\sqrt{w} = 0. \qquad [3]$$

Esto implica—y aquí está la clave del argumento—que a y b son iguales a cero. Pues si b no fuera cero, podríamos deducir de [3] que $\sqrt{w} = -a/b$, y $\sqrt{w}$ sería un número del cuerpo F_{k-1} al que pertenecen a y b, contrariamente a nuestra hipótesis. Siendo, pues, $b = 0$, se sigue inmediatamente de [3] que también $a = 0$.

Una vez visto que $a = b = 0$, es consecuencia inmediata de [2'] que $y = p - q\sqrt{w}$ es también solución de la ecuación cúbica [1], por ser $y^3 - 2 = 0$. Además, $y \neq x$; es decir, $x - y \neq 0$, pues $x - y = 2q\sqrt{w}$ sólo puede anularse si $q = 0$, de donde $x = p$ pertenecería a F_{k-1}, contra lo supuesto.

Hemos probado así que si $x = p + q\sqrt{w}$ es una solución de la ecuación cúbica [1], también $y = p - q\sqrt{w}$ es otra solución diferente de esta ecuación, lo que entraña una evidente contradicción, pues hay un único número real x que es raíz cúbica de 2, siendo las otras dos raíces cúbicas imaginarias (véase pág. 107); $y = p - q\sqrt{w}$ es, sin embargo, real, ya que p, q, y $\sqrt{w}$ son reales.

La hipótesis hecha nos ha llevado a un absurdo, quedando así demostrada su falsedad. Una solución de [1] no puede estar en el cuerpo F_k; es decir, es imposible duplicar el cubo con regla y compás.

2. Un teorema sobre ecuaciones cúbicas.—El razonamiento algebraico que acabamos de utilizar ha sido adaptado especialmente al problema particular analizado. Si deseamos ocuparnos de los otros dos problemas clásicos, es preferible proceder sobre una base más general. Los tres problemas dependen algébricamente de ecuaciones cúbicas. Un hecho fundamental concerniente a la ecuación cúbica

$$z^3 + az^2 + bz + c = 0 \qquad [4]$$

es la relación siguiente entre las tres raíces x_1, x_2, x_3 de la misma:

$$x_1 + x_2 + x_3 = -a\ *. \qquad [5]$$

Consideremos la ecuación cúbica [4] con coeficientes a, b, c, racionales. Puede suceder que una de las raíces de la ecuación sea racional; p. ej., la ecuación $x^3 - 1 = 0$ tiene la raíz racional 1, mientras que las otras dos raíces, dadas por la ecuación cuadrática $x^2 + x + 1 = 0$, son necesariamente imaginarias. Pero podemos demostrar fácilmente el siguiente teorema general: *Si una ecuación cúbica de coeficientes racionales no tiene raíz racional, ninguna de sus raíces es construíble partiendo del cuerpo racional* F_0.

Daremos nuevamente una demostración por reducción al absurdo. Supongamos que x fuera una raíz construíble de [4]. Entonces x pertenecería al último cuerpo F_k de alguna cadena de cuerpos sucesivamente ampliados $F_0, F_1 \ldots F_k$, como antes. Supondremos que k es el *menor* entero tal que una raíz de la ecuación cúbica [4] pertenece a F_k. Por supuesto que k debe ser mayor que cero, ya que en el enunciado del teorema se supone que ninguna raíz x pertenece al cuerpo racional F_0; luego x puede escribirse en la forma

$$x = p + q\sqrt{w},$$

donde p, q, w pertenecen al cuerpo precedente F_{k-1}, pero $\sqrt{w}$ no. Sigue de ello, exactamente como para la ecuación particular $z^3-2=0$ anterior, que otro número de F_k,

$$y = p - q\sqrt{w},$$

será también solución de la ecuación [4]. Como antes, podremos ver que $q \neq 0$ y, por tanto, $x \neq y$.

De [5] resulta que la tercera raíz de la ecuación [4] está dada por $u = -a - x - y$. Pero, dado que $x + y = 2p$, esto significa que

$$u = -a - 2p,$$

y como $\sqrt{w}$ ha desaparecido, u es un número del cuerpo F_{k-1}. Esto contradice la hipótesis de que k es el *menor* número tal que algún F_k contiene a una raíz de [4]; en consecuencia, la hipótesis es absurda, y

* El polinomio $z^3 + az^2 + bz + c$ puede descomponerse en el producto $(z - x_1)(z - x_2)(z - x_3)$, siendo x_1, x_2, x_3 las raíces de la ecuación [4] (véase pág. 110). Por tanto,

$$z^3 + az^2 + bz + c = z^3 - (x_1 + x_2 + x_3)z^2 + (x_1x_2 + x_1x_3 + x_2x_3)z - x_1x_2x_3,$$

y como el coeficiente de cada potencia debe ser el mismo en ambos miembros, resulta:

$$-a = x_1 + x_2 + x_3, \qquad b = x_1x_2 + x_1x_3 + x_2x_3, \qquad -c = x_1x_2x_3.$$

ninguna raíz de [4] pertenece a ningún cuerpo F_k. El teorema general está demostrado. Basándose en este teorema quedará probada la imposibilidad de una construcción con regla y compás si el equivalente algebraico del problema resulta ser solución de una ecuación cúbica desprovista de raíces racionales. Esta equivalencia es obvia para el problema de duplicar el cubo, y vamos a establecerla para los otros dos problemas griegos.

3. Trisección del ángulo.—Vamos a demostrar que la trisección del ángulo con la regla y el compás es, *en general*, imposible. Naturalmente, existen ángulos como los de 90° y 180°, para los cuales es posible la trisección. Lo que vamos a demostrar es que la trisección no puede efectuarse por un método válido para *todo* ángulo. Para ello basta considerar un ángulo que no pueda trisecarse, ya que un *método general* debería ser válido para cualquier ejemplo. Por consiguiente, la no existencia de un método general puede probarse si se demuestra que el ángulo de 60°, p. ej., no puede ser trisecado sólo con el auxilio de la regla y el compás.

Podemos obtener un equivalente algebraico de este problema de diferentes formas; la más sencilla consiste en considerar el ángulo por su coseno: $\cos\theta = g$. Entonces, el problema es el de encontrar la cantidad $x = \cos(\theta/3)$. Mediante una sencilla fórmula trigonométrica (véase pág. 106) el coseno de $\theta/3$ se halla ligado con el de θ por la ecuación

$$\cos\theta = g = 4\cos^3(\theta/3) - 3\cos(\theta/3).$$

En otras palabras, el problema de trisecar el ángulo θ, con $\cos\theta = g$, equivale a construir una solución de la ecuación cúbica

$$4z^3 - 3z - g = 0. \qquad [6]$$

Para probar que en general no puede hacerse esto, tomemos $\theta = 60°$, de donde $g = \cos 60° = 1/2$, y la ecuación [6] se convierte en

$$8z^3 - 6z = 1. \qquad [7]$$

En virtud del teorema antes demostrado, basta probar que esta ecuación no tiene raíz racional. Haciendo $v = 2z$, la ecuación se transforma en

$$v^3 - 3v = 1. \qquad [8]$$

Si hubiera un número racional $v = r/s$ que verificara esta ecuación (r y s primos entre sí), tendríamos $r^3 - 3s^2r = s^3$; de donde $s^3 = r(r^2 - 3s^2)$ sería divisible por r, y r y s tendrían un factor común,

a menos que $r = \pm 1$. Asimismo, s^2 divide a $r^3 = s^2(s + 3r)$, lo que supone que r y s tienen algún factor común, salvo que $s = \pm 1$. Como hemos supuesto que s y r carecen de factor común, vemos que los únicos números racionales que pueden verificar la ecuación [8] son $+1$ ó -1. Pero sustituyendo $+1$ y -1 en lugar de v en [8] vemos que ninguno de ambos la satisface; luego [8] y, en consecuencia, [7] carecen de raíz racional, y la imposibilidad de trisecar el ángulo queda demostrada.

Este teorema de que un ángulo no puede ser trisecado con regla y compás sólo es cierto cuando la regla se usa *exclusivamente* como instrumento para trazar la recta determinada por dos puntos. En nuestra caracterización general de los números construíbles, el uso de la regla se ha limitado siempre a esta operación; si se permiten otros usos de la regla, la totalidad de las construcciones posibles puede extenderse enormemente. El siguiente método para trisecar el ángulo, utilizado en las obras de Arquímedes, es un buen ejemplo. Sea x un ángulo arbitrario dado, como en la figura 36. Prolonguemos la base del ángulo por la izquierda y tracemos un semicírculo con centro en 0 y radio arbitrario r. Señalemos dos puntos A y B en el borde de la regla, tales que $AB = r$. Manteniendo el punto B en la semicircunferencia, deslicemos la regla haciendo que A caiga sobre la prolongación del lado inicial del ángulo x, mientras el borde de la regla pasa por la intersección C del segundo lado del ángulo x con la semicircunferencia de centro O. Con la regla en esta posición tracemos una recta, que forma un ángulo y con la prolongación del lado inicial del ángulo x.

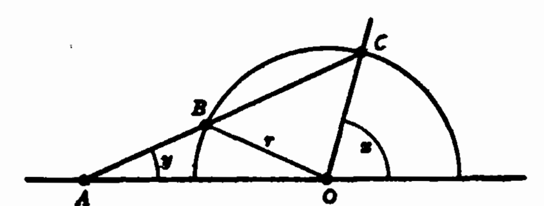


FIG. 36.—Trisección de un ángulo (Arquímedes).

Ejercicio: Demuéstrese que esta construcción nos da $y = x/3$.

4. El heptágono regular.—Vamos a considerar ahora el problema de hallar el lado z de un heptágono regular inscrito en la circunferencia unidad. La forma más sencilla de tratar este problema es utilizar los números complejos (Cap. II, V); sabemos que los vértices del heptágono están dados por las raíces de la ecuación

$$z^7 - 1 = 0, \qquad [9]$$

siendo las coordenadas x, y de los vértices las partes real e imaginaria

del número complejo $z = x + yi$. Una solución de esta ecuación es $z = 1$, y las otras son las raíces de la ecuación

$$\frac{z^7 - 1}{z - 1} = z^6 + z^5 + z^4 + z^3 + z^2 + z + 1 = 0, \qquad [10]$$

obtenida de [9] dividiendo por el factor $z - 1$ (pág. 108). Dividiendo [10] por z^3, obtenemos la ecuación

$$z^3 + 1/z^3 + z^2 + 1/z^2 + z + 1/z + 1 = 0. \qquad [11]$$

Mediante una sencilla transformación algebraica, podemos escribir [11] en la forma:

$$(z + 1/z)^3 - 3(z + 1/z) + (z + 1/z)^2 - 2 + (z + 1/z) + 1 = 0. \qquad [12]$$

Designando $z + 1/z$ por y, de [12] deducimos:

$$y^3 + y^2 - 2y - 1 = 0. \qquad [13]$$

Sabemos que z, raíz séptima de la unidad, está dada por

$$z = \cos \varphi + i \operatorname{sen} \varphi, \qquad [14]$$

donde $\varphi = 360^\circ/7$ es el ángulo central subtendido por el lado del heptágono regular; asimismo sabemos (Ej. 2, pág. 106) que $1/z = \cos \varphi - i \operatorname{sen} \varphi$; es decir, $y = z + 1/z = 2 \cos \varphi$.

Si es posible construir y, podremos también construir $\cos \varphi$, y recíprocamente. Luego si podemos probar que y no es construíble, quedará al mismo tiempo probado que z, y, por tanto, el heptágono, no es construíble. De esta forma, en virtud del teorema antes demostrado, queda solamente por probar que la ecuación [13] no tiene raíces racionales. Esto también será probado por reducción al absurdo. Supongamos que [13] tiene una raíz racional r/s (r y s primos entre sí). Tenemos

$$r^3 + r^2s - 2rs^2 - s^3 = 0; \qquad [15]$$

donde se ve como antes que r^3 es divisible por s, y s^3 por r. Como s y r son primos entre sí, cada uno debe ser igual a ± 1; por tanto, si y es racional, sólo puede tomar los valores $+1$ y -1. Sustituyendo estos números en la ecuación, vemos que ninguno de ellos la satisface. Luego y, y por ende el heptágono regular, no es construíble.

5. Observaciones acerca de la cuadratura del círculo.—Hemos logrado tratar estos problemas de duplicar el cubo, trisecar el ángulo y construir el heptágono regular, mediante métodos relativamente

elementales. El problema de la cuadratura del círculo es mucho más difícil y requiere la técnica del análisis matemático superior. Como un círculo de radio r tiene área πr^2, el problema de construir un cuadrado de área igual a la de un círculo de radio unidad equivale a la construcción de un segmento de longitud $\sqrt{\pi}$, lado del cuadrado pedido. Este segmento será construíble si, y sólo si, el número π es construíble. En virtud de nuestra caracterización general de los números construíbles, podemos demostrar la imposibilidad de cuadrar el círculo probando que el número π no puede estar contenido en ningún cuerpo F_k que pueda deducirse del cuerpo racional F_0 mediante sucesivas adjunciones de raíces cuadradas. Como todos los elementos de tal cuerpo son números algebraicos, es decir, números que satisfacen a ecuaciones algebraicas de coeficientes enteros, nos basta con probar que el número π no es algebraico; es decir, que es *trascendente* (véase pág. 112).

La técnica necesaria para establecer que π es un número trascendente fué creada por Charles Hermite (1822-1905), quien demostró que el número e es trascendente. Mediante un ligero perfeccionamiento del método de Hermite, F. Lindemann consiguió (1882) probar la trascendencia de π, poniendo fin para siempre a la vieja cuestión de la cuadratura del círculo. La demostración está al alcance del estudiante de análisis superior, pero excede los fines de este libro.

PARTE SEGUNDA

VARIOS MÉTODOS PARA OBTENER CONSTRUCCIONES

IV. TRANSFORMACIONES GEOMÉTRICAS. INVERSIÓN

1. Observaciones generales.—En la segunda parte de este capítulo vamos a discutir de forma sistemática algunos principios generales que pueden aplicarse a los problemas de construcción. Muchos de estos problemas pueden dominarse con más claridad desde el punto de vista general de las «transformaciones geométricas»; en lugar de estudiar una construcción particular, vamos a considerar simultáneamente la totalidad de los problemas ligados por ciertos procesos de transformación. El poder de síntesis aclaratoria del concepto de clase de transformaciones geométricas no está en modo alguno restringido a los problemas de construcción, sino que afecta a casi toda la geometría. En los capítulos IV y V nos ocuparemos de este aspecto general de las transformaciones geométricas, limitándonos ahora a estudiar un tipo particular de transformación: la inversión respecto a una circunferencia del plano, que es una generalización de la simetría ordinaria respecto a una recta.

Por *transformación* o *representación* del plano en sí mismo entendemos una ley que asigna a cada punto P del plano otro punto P', llamado *imagen* de P en la transformación; el punto P se llama *antecedente* de P'. Un ejemplo sencillo de tal transformación es la simetría del plano respecto de una recta L, como en un espejo; un punto P, situado a un lado de L, tiene como imagen el punto P' del otro lado de L, y tal que L es la mediatriz del segmento PP'. Una transformación puede dejar fijos ciertos puntos del plano; en el caso de la simetría, esto ocurre para los puntos de L.

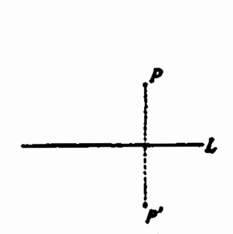


FIG. 37.—Simetría de un punto respecto a una recta.

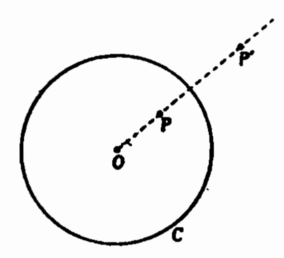


FIG. 38.—Inversión de un punto respecto a una circunferencia.

Otros ejemplos de transformaciones son: las *rotaciones* del plano alrededor de un punto fijo O; las *traslaciones* paralelas, que trasladan cada punto una distancia d en una dirección dada (tales transformaciones no dejan puntos fijos); y, más en general, los *movimientos rígidos* del plano, que pueden imaginarse como compuestos de rotaciones y traslaciones paralelas.

La clase particular de transformaciones que ahora nos interesa es la de las *inversiones* respecto a circunferencias. (Algunas veces llamadas *reflexiones circulares*, debido a que representan con cierta aproximación la relación entre el objeto y la imagen en una reflexión sobre un espejo circular.) En un plano dado, sea C una circunferencia de centro O (llamado centro de la inversión) y radio r. Definimos como imagen del punto P, el punto P' de la recta OP, situado del mismo lado de O que P, y tal que cumple la condición

$$OP \cdot OP' = r^2. \qquad [1]$$

Los puntos P y P' se denominan *puntos inversos* respecto a C. De esta definición concluímos que si P' es el punto inverso de P, a su vez P es el inverso de P'. Una inversión intercambia el interior y el exterior del círculo C, ya que para $OP < r$ tenemos $OP' > r$, y para $OP > r$, $OP' < r$. Los únicos puntos del plano que quedan fijos en la inversión son los puntos de la circunferencia C.

La regla [1] no define una imagen para el centro O. Es evidente que si un punto P móvil lo aproximamos a O, la imagen P' se aleja cada vez más en el plano. Por esta razón decimos a veces que O corresponde al *punto del infinito* en la inversión. La utilidad de esta forma de hablar reside en el hecho de permitirnos asegurar que una inversión establece una correspondencia entre los puntos del plano y sus imágenes, que es biunívoca sin excepción: cada punto del plano tiene una imagen y sólo una, y él mismo es imagen de un punto y sólo uno. Todas las transformaciones consideradas anteriormente gozan de esta propiedad.

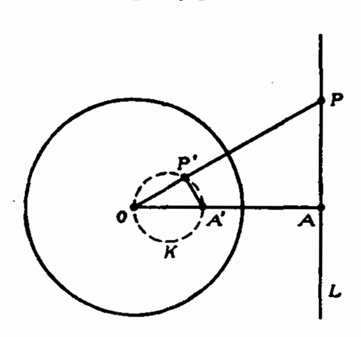


FIG. 39.—Inversión de una recta L respecto a una circunferencia.

2. Propiedades de la inversión.—La propiedad más importante de la inversión es la de que transforma rectas y circunferencias en rectas y circunferencias. Con más precisión, vamos a ver que en una inversión:

a) Una recta que pasa por O se transforma en una recta que pasa por O.

b) Una recta que no pasa por O se transforma en una circunferencia que pasa por O.

c) Una circunferencia que pasa por O se transforma en una recta que no pasa por O.

d) Una circunferencia que no pasa por O se transforma en una circunferencia que no pasa por O.

La proposición *a)* es obvia, ya que por definición de inversión todo punto de la recta tiene como imagen otro punto de la misma; es decir, que aunque los puntos de la recta se intercambian, la recta como totalidad se transforma en sí misma.

Para probar *b)* tracemos una perpendicular desde O a la recta L (Fig. 39). Sea A el punto donde esta perpendicular corta a L, y A'

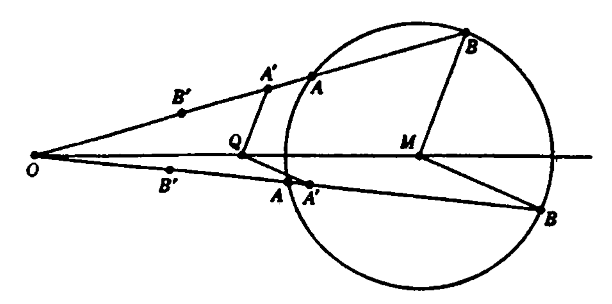


FIG. 40.—Inversión de una circunferencia.

el inverso de A. Tomemos un punto cualquiera P de L, y sea P' su inverso. Como $OA' \cdot OA = OP' \cdot OP = r^2$, resulta que

$$OA'/OP' = OP/OA.$$

Por tanto, los triángulos $OP'A'$ y OAP son semejantes y el ángulo $OP'A'$ es recto. Por geometría elemental sabemos que P' está en la circunferencia K de diámetro OA', de donde la inversa de L es esta circunferencia. Esto prueba *b)*. La proposición *c)* se demuestra por el hecho de que si K es la inversa de L, la inversa de K es L.

Queda por demostrar la proposición *d)*. Sea K una circunferencia que no pasa por O, de centro M y radio k. Para obtener su imagen, tracemos una recta por O que corte a K en A y B, y veamos cómo varían las imágenes A' y B' cuando la recta que pasa por O corta a K de todas las formas posibles. Designemos las distancias OA, OB, OA', OB', OM por a, b, a', b', m, y sea t la longitud de la tangente

a K desde O. Tenemos $aa' = bb' = r^2$, por definición de inversión, y $ab = t^2$, por una propiedad geométrica elemental del círculo. Si dividimos las primeras relaciones por la segunda, resulta

$$a'/b = b'/a = r^2/t^2 = c^2,$$

donde c^2 es una constante que depende sólo de r y t, y es la misma para todas las posiciones de A y B. Por A' trazamos una paralela a BM que corta a OM en Q. Sea $OQ = q$, y $A'Q = \rho$; entonces $q/m = a'/b = \rho/k$, o sea

$$q = ma'/b = mc^2, \qquad \rho = ka'/b = kc^2.$$

Esto significa que para todas las posiciones de A y B, Q será siempre el mismo punto de OM, y la distancia $A'Q$ tendrá siempre el mismo valor. Además, $B'Q = \rho$, ya que $a'/b = b'/a$. Así, las imágenes de todos los puntos A, B, de K, son puntos cuyas distancias a Q son iguales a ρ; es decir, la imagen de K es una circunferencia. Esto prueba *d)*.

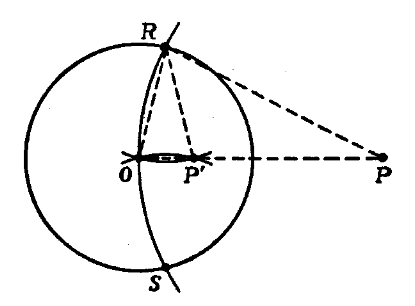


Fig. 41.—Inversión de un punto exterior respecto a una circunferencia.

3. Construcción geométrica de puntos inversos.—El siguiente teorema nos será útil en el párrafo próximo: *El punto* P', *inverso de un punto dado* P *respecto a una circunferencia* C, *puede ser construído geométricamente mediante el solo uso del compás.* Consideremos primero el caso en que el punto dado P sea exterior a C. Con OP como radio y centro en P describimos un arco que corte a C en los puntos R y S. Con estos dos puntos como centros describimos arcos de radio r que se cortan en O y en el punto P' de la recta OP. En los triángulos isósceles ORP y ORP' se verifica

$$\widehat{ORP} = \widehat{POR} = \widehat{OP'R},$$

luego estos triángulos son semejantes, y se tiene:

$$\frac{OP}{OR} = \frac{OR}{OP'}; \quad \text{esto es,} \quad OP \cdot OP' = r^2.$$

El punto P' así construído es, por tanto, el inverso de P.

Si el punto dado P es interior a C, subsiste la misma construcción, siempre que la circunferencia de radio OP y centro P corte a C en

dos puntos. Si no la corta, podemos reducir la construcción del punto inverso P' al caso anterior mediante el siguiente artificio:

Observemos primero que con el solo uso del compás podemos encontrar un punto C de la recta que une dos puntos dados A, O, y tal que $AO = OC$. Para esto, tracemos una circunferencia de centro O y radio $r = OA$, y llevemos sobre esta circunferencia a partir de A, los puntos P, Q, C, tales que $AP = PQ = QC = r$. Entonces C es el punto deseado, como se ve por el hecho de que los triángulos AOP, OPQ, OQC son equiláteros; es decir, OA y OC forman un ángulo de 180º, y $OC = OQ = AO$. Repitiendo este procedimiento, podemos fácilmente prolongar AO un número deseado de veces. Incidentalmente, como la longitud del segmento AQ es $r\sqrt{3}$, como el lector puede verificar fácilmente, hemos construído al mismo tiempo $\sqrt{3}$, a partir de la unidad, sin utilizar la regla.

Podemos ahora encontrar el inverso de un punto P interior a la circunferencia C. Primero hallaremos un punto R de la recta OP cuya distancia a O sea un múltiplo entero de OP y que quede exterior a C; es decir,

$$OR = n \cdot OP.$$

Podemos hacer esto llevando sucesivamente la distancia OP con el compás hasta salir fuera de C. Hallamos ahora el punto R' inverso del R, mediante la construcción antes dada. Entonces,

$$r^2 = OR' \cdot OR = OR' \cdot (n \cdot OP) = (n \cdot OR') \cdot OP.$$

Por tanto, P', tal que $OP' = n \cdot OR'$, es el punto inverso buscado.

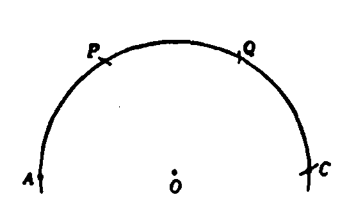


FIG. 42.—Duplicación de un segmento.

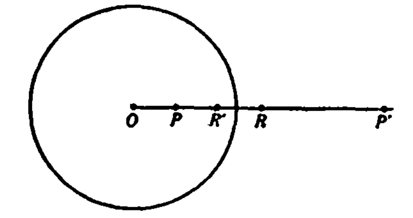


FIG. 43.—Inversión de un punto interior respecto a una circunferencia.

4. Forma de hallar sólo con el compás el punto medio de un segmento y el centro de una circunferencia.—Habiendo aprendido ya a determinar el inverso de un punto dado mediante el uso exclusivo del

compás, podemos realizar algunas construcciones interesantes; p. ej., consideremos el problema de encontrar el punto medio entre dos puntos A y B, utilizando solamente el compás (¡sin trazar rectas!). He aquí la solución: tracemos la circunferencia de radio AB y centro B, y llevemos tres arcos de radio AB, a partir de A. El punto final C estará en la recta AB, siendo $AB = BC$. Dibujemos ahora la circunferencia de radio AB y centro A, y sea C' el punto inverso de C respecto a este círculo. Entonces: $AC' \cdot AC = AB^2$; o sea, $AC' \cdot 2AB = AB^2$, y $2AC' = AB$; por tanto, C' es el punto medio pedido.

Otra construcción con compás, que hace uso de puntos inversos, es la de hallar el centro de una circunferencia dada. Elegimos un punto P de la circunferencia y con centro en él trazamos otra circun-

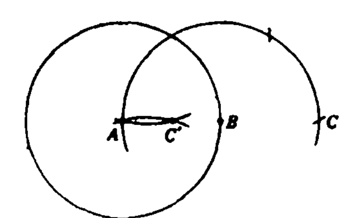


FIG. 44.—Determinación del punto medio de un segmento.

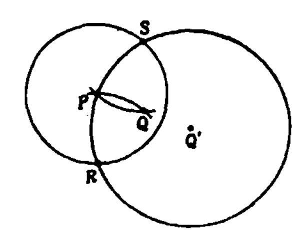


FIG. 45.—Determinación del centro de un círculo.

ferencia que corte a la dada en los puntos R y S. Con estos centros trazamos arcos de radios $RP = SP$, que se cortan en Q. Comparando con la figura 41, vemos que el centro desconocido Q' es inverso de Q respecto al círculo de centro P, por lo que Q' puede ser construído haciendo sólo uso del compás.

V. CONSTRUCCIONES CON OTROS INSTRUMENTOS. CONSTRUCCIONES DE MASCHERONI CON COMPÁS SOLAMENTE

***1. Una construcción clásica para duplicar el cubo.**—Hasta aquí hemos considerado únicamente problemas de construcciones geométricas con regla y compás. Cuando se permiten otros instrumentos, la variedad de las construcciones posibles se hace naturalmente más extensa; p. ej., los griegos resolvían el problema de duplicar el cubo de la forma siguiente: consideremos (Fig. 46) un ángulo recto rígido MZN y una cruz movible en ángulo recto B, VW, PQ. Dos varillas adicionales RS y TU pueden deslizarse perpendicularmente a los lados del ángulo recto. En la cruz se eligen dos puntos fijos E y G, de tal forma que $GB = a$ y $BE = f$ tengan longitudes prefijadas. Situando

la cruz de modo que los puntos E y G estén sobre NZ y MZ, respectivamente, y deslizando las varillas TU y RS, podemos llevar el aparato entero a una posición en que tengamos el rectángulo $ADEZ$, por cuyos vértices A, D, E, pasan los brazos BW, BQ, BV de la cruz. Tal disposición es siempre posible, si $f > a$. Vemos entonces que $a : x = x : y = y : f$; de donde, si f es igual a $2a$ en el aparato, resulta

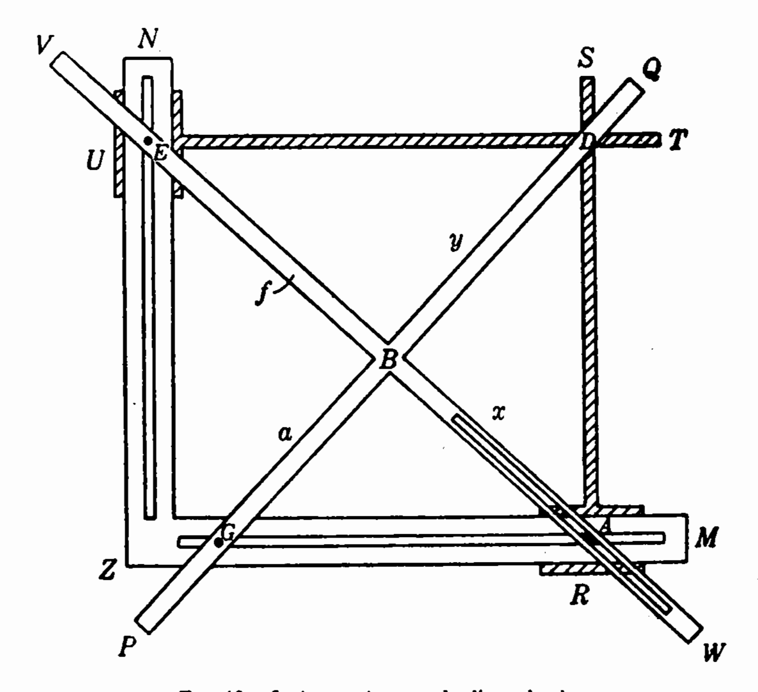


FIG. 46.—Instrumento para duplicar el cubo.

$x^3 = 2a^3$; por tanto, x es la arista de un cubo cuyo volumen es doble del cubo de arista a. Tenemos así resuelto el problema de duplicar el cubo.

2. Restricción de usar sólo el compás.—Resulta natural que si se permite una mayor variedad de instrumentos podamos resolver una colección más amplia de problemas de construcción, y cabe esperar que a una mayor restricción de instrumentos corresponda una clase más restringida también de las construcciones posibles; por ello, fué un descubrimiento sorprendente, debido al italiano Mascheroni (1750-1800), el de que *todas las construcciones geométricas posibles mediante la regla y el compás pueden hacerse sólo con el compás*. Naturalmente, no se puede trazar la recta que une dos puntos sin la regla, por lo

que esta construcción fundamental no está en realidad comprendida en la teoría de Mascheroni. En cambio, puede suponerse que la recta está dada por dos de sus puntos. Mediante el uso del compás solo, se puede encontrar el punto de intersección de dos rectas dadas de este modo, como también las intersecciones de una circunferencia dada con una recta.

Quizá el ejemplo más sencillo de construcción de Mascheroni sea el de duplicar un segmento dado AB. La solución fué dada en la página 157, y en la siguiente hemos determinado el punto medio de un segmento. Vamos ahora a resolver el problema de bisecar un arco AB de una circunferencia de centro dado O. La construcción es la siguiente: desde A y B como centros trazamos dos arcos, de radio AO; a partir de O, llevamos los arcos OP y OQ iguales a AB. A continuación, dibujamos dos arcos de radios PB y QA y centros P y Q, que se cortan en R. Finalmente, con radio OR, describimos un arco con centro en P o Q hasta cortar a AB; el punto de intersección es el punto medio buscado del arco AB. Dejamos la demostración como ejercicio al lector.

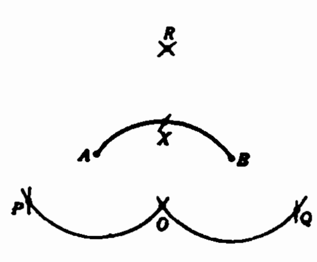


FIG. 47.—Bisección de un arco con el compás.

Sería imposible demostrar el teorema general de Mascheroni dando la construcción con compás solo para cada construcción posible con regla y compás, ya que el número de éstas no es finito. Sin embargo, podemos lograr el mismo objeto demostrando que cada una de las cuatro construcciones fundamentales siguientes es posible con el compás solo;

1) Trazar una circunferencia de centro y radio dados.

2) Hallar los puntos de intersección de dos circunferencias.

3) Hallar los puntos de intersección de una recta y una circunferencia.

4) Hallar el punto de intersección de dos rectas.

Toda construcción geométrica en el sentido usual, cuando se permite el uso de regla y compás, consiste en una sucesión finita de estas construcciones elementales. Las dos primeras son evidentemente posibles con compás solo. Las soluciones de los problemas más difíciles 3 y 4 dependen de las propiedades de la inversión desarrollada en la sección precedente.

Resolvamos el problema 3, consistente en hallar los puntos de intersección de una circunferencia C y una recta, dada por los puntos

A y *B*. Con centros *A* y *B* y radios *AO* y *BO*, respectivamente, dibujemos dos arcos, que se cortarán de nuevo en *P*. Determinemos ahora el punto *Q* inverso de *P* respecto a *C*, mediante la construcción con compás solo dada en la página 157. Tracemos la circunferencia de centro *Q* y radio *QO* (que debe cortar a *C*); los puntos de intersección *X* y *X'* de esta circunferencia con la dada, *C*, son los dos puntos buscados. Para probarlo necesitamos solamente demostrar que *X* y *X'* son equidistantes de *O* y *P*, ya que *A* y *B* lo son por construcción. Esto resulta del hecho de que el inverso de *Q* es un punto cuya distancia a *X* y *X'* es igual al radio de *C* (pág. 157). Observemos que la circun-

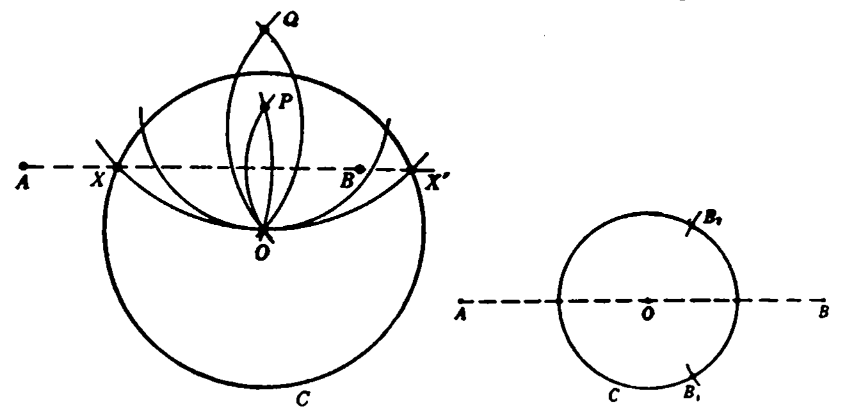


Fig. 48.—Intersección de una circunferencia con una recta que no pasa por su centro.

Fig. 49.—Intersección de una circunferencia con una recta que pasa por su centro.

ferencia que pasa por *X*, *X'* y *O* es inversa de la recta *AB*, ya que ésta y la circunferencia deben cortar a *C* en los mismos puntos. (Los puntos de la circunferencia son inversos de sí mismos.)

La construcción deja de ser válida sólo en el caso de que la recta *AB* pase por el centro de *C*. Pero entonces los puntos de intersección pueden determinarse mediante la construcción dada en la página 160, como puntos medios de arcos de *C*, obtenidos trazando con centro en *B* una circunferencia arbitraria que corte a *C* en B_1 y B_2.

El método para determinar la circunferencia inversa de la recta que une dos puntos dados permite una solución inmediata del problema 4. Sean dos rectas dadas *AB* y *A'B'* (Fig. 50). Tracemos una circunferencia cualquiera *C* del plano y, mediante el método precedente, hallemos las circunferencias inversas de *AB* y *A'B'*, las cuales se cortan en *O* y en el punto *Y*. El punto *X*, inverso del *Y*, es el punto pedido, y puede construirse mediante el procedimiento ya dado. Que

X es el punto buscado es evidente por el hecho de que Y es el único punto con la propiedad de ser a la vez inverso de un punto de AB y de $A'B'$; luego el punto X, inverso del Y, debe estar en AB y $A'B'$.

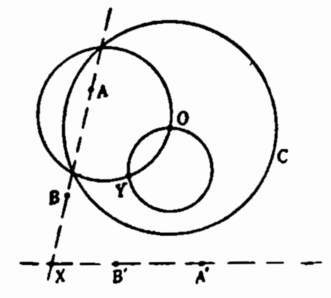


FIG. 50.—Intersección de dos rectas.

Con estas dos construcciones hemos completado la demostración de equivalencia entre las construcciones de Mascheroni sólo con el compás, y las construcciones geométricas usuales con regla y compás. No nos hemos cuidado de dar soluciones elegantes para cada problema particular, ya que nuestro objeto ha sido más bien el de penetrar en el fondo general de las construcciones de Mascheroni. Vamos a dar como ejemplo, sin embargo, las construcción del pentágono regular; con más precisión, vamos a hallar cinco puntos de una circunferencia que sean vértices de un pentágono regular inscrito.

Sea A un punto de la circunferencia dada K. El lado del hexágono regular inscrito es igual al radio de K, por lo que podemos hallar los puntos B, C, D sobre K, tales que $\widehat{AB} = \widehat{BC} = \widehat{CD} = 60^\circ$ (figura 51). Con A y D como centros y radio AC, tracemos dos arcos que se cortarán en X. Entonces, si es O el centro de K, un arco de centro A y radio OX cortará a K en el punto medio F de $\widehat{BC}$ (véase pág. 160). Ahora, con el radio de K, tracemos un arco con centro en F, que encuentra a K en G y H. Sea Y un punto cuya distancia a G y H es OX, y que está separado de X por O. Entonces AY será igual al lado del pentágono buscado. La

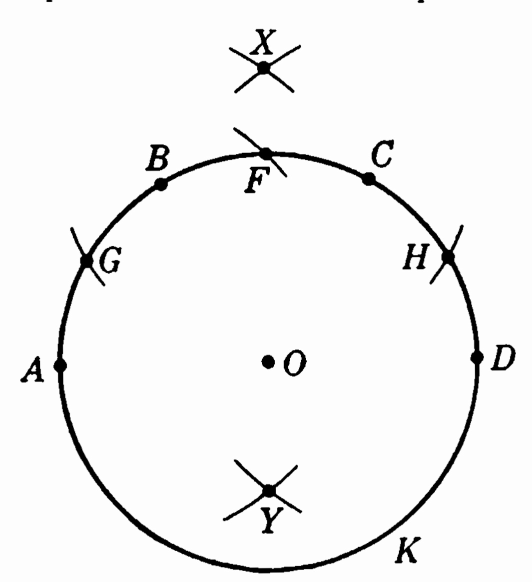


FIG. 51.—Construcción del pentágono regular.

demostración se deja como ejercicio al lector. Observemos que se han utilizado en la construcción solamente tres radios diferentes.

En 1928, el matemático danés Hjelmslev halló, en una librería de Copenhague, un libro llamado *Euclides Danicus*, publicado en 1672 por un oscuro autor, G. Mohr. Del título podría inferirse que este trabajo era simplemente una versión o comentario de los *Elementos* de Euclides. Pero cuando Hjelmslev examinó el libro se sorprendió al encontrar que contenía esencialmente el problema de Mascheroni y su solución completa, encontrada antes que Mascheroni.

Ejercicios: He aquí la descripción de las construcciones de Mohr. Confróntese su validez. ¿Por qué resuelven el problema de Mascheroni?

1. Sobre un segmento AB de longitud p trácese un segmento perpendicular BC. (*Solución:* prolónguese AB hasta un punto D, tal que $AB = BD$. Trácense circunferencias arbitrarias con centros A y D, determinando así C.)
2. Dados dos segmentos de longitudes p y q tales que $p > q$, hállese un segmento de longitud $x = \sqrt{p^2 - q^2}$, haciendo uso de 1.
3. Dado un segmento a, constrúyase el segmento $a\sqrt{2}$. [Obsérvese que $(a\sqrt{2})^2 = (a\sqrt{3})^2 - a^2$.]
4. Con dos segmentos dados, p y q, hállese un segmento $x = \sqrt{p^2 + q^2}$. [Utilícese la relación $x^2 = 2p^2 - (p^2 - q^2)$.] Háganse otras construcciones similares.
5. Utilizando los resultados anteriores, constrúyanse segmentos de longitudes $p + q$ y $p - q$, supuestos dados en el plano los segmentos de longitudes p y q.
6. Compruébese y demuéstrese la siguiente construcción para el punto medio M de un segmento dado AB de longitud a. En la prolongación de AB constrúyanse C y D tales que $CA = AB = BD$. Trácese el triángulo isósceles ECD con $EC = ED = 2a$, y hállese M como intersección de dos circunferencias de diámetros EC y ED.
7. Hállese la proyección ortogonal de un punto A sobre la recta BC.
8. Constrúyase x tal que $x : a = p : q$, si a, p, q son segmentos dados.
9. Hállese $x = ab$, si a y b son segmentos dados.

Inspirado por Mascheroni, Jacob Steiner (1796-1863) trató de utilizar sólo como instrumento la regla en lugar del compás. Naturalmente, la regla sola no puede llevarnos fuera de un cuerpo de números dado, por lo cual no basta para todos las construcciones geométricas en el sentido clásico. Lo notable es que Steiner fué capaz de restringir el uso del compás a una sola aplicación. Demostró que todas las construcciones en el plano que son posibles con regla y compás pueden hacerse sólo con la regla, siempre que se suponga dada *una* circunferencia fija, y su centro. Estas construcciones requieren métodos proyectivos y serán indicadas más adelante (véase Cap. IV, pág. 209).

* Esta circunferencia y su centro son imprescindibles; p. ej., si se da un círculo, pero no su centro, es imposible construir este último mediante la regla sola. Para demostrarlo hay que hacer uso de un resultado que será discutido más adelante

(Cap. IV, pág. 232). Existe una transformación del plano en sí mismo que tiene las siguientes propiedades: *a*) el círculo dado queda fijo en la transformación; *b*) toda recta se transforma en recta; *c*) el centro del círculo se transforma en algún otro punto. La mera existencia de tal transformación prueba la imposibilidad de contruir con la regla sola el centro de un círculo dado. Pues si la construcción fuera posible, consistiría en dibujar un cierto número de rectas y determinar sus intersecciones, entre sí, y con la circunferencia dada. Ahora bien: si a la figura total, formada por la circunferencia dada y todos los puntos y rectas de la construcción, se le aplica la transformación cuya existencia hemos supuesto, la figura transformada satisfará todas las exigencias de la construcción, pero dará como resultado un punto distinto del centro del círculo. Por tanto, tal construcción es imposible.

3. Trazado con instrumentos mecánicos. Curvas mecánicas. Cicloides.—Ideando mecanismos para dibujar curvas distintas de la circunferencia y de la recta, podemos ampliar el dominio de las figuras construíbles; p. ej., si tenemos un instrumento para trazar las hipérbolas $xy = k$, y otro para dibujar las parábolas $y = ax^2 + bx + c$, entonces todo problema que conduzca a la ecuación cúbica

$$ax^3 + bx^2 + cx = k, \qquad [1]$$

puede ser resuelto por construcción, utilizando sólo estos instrumentos. Pues si hacemos

$$xy = k, \qquad y = ax^2 + bx + c, \qquad [2]$$

resolver [1] equivale a resolver el sistema de ecuaciones [2] por eliminación de y; es decir, las soluciones de [1] son las abscisas x de los puntos de intersección de la hipérbola y la parábola de [2]. Así, las soluciones de [1] podrían construirse si dispusiéramos de instrumentos para dibujar la parábola y la hipérbola de las ecuaciones [2].

Desde la antigüedad, los matemáticos sabían que muchas curvas interesantes podían ser definidas y dibujadas mediante sencillos instrumentos mecánicos. De estas «curvas mecánicas» las *cicloides* son de las más notables. Tolomeo (200 a. de J. C.) las utilizó en forma muy ingeniosa para describir los movimientos de los planetas del sistema solar.

La cicloide más sencilla es la curva descrita por un punto fijo de la circunferencia de un círculo cuando rueda sin deslizar sobre una recta. La figura 53 nos muestra cuatro posiciones del punto P del círculo rodante. El aspecto general de la cicloide es el de una serie de arcos que se apoyan sobre la recta. Pueden obtenerse variaciones de esta curva eligiendo el punto P, ya en el interior del círculo (como en

FIG. 52.—Solución gráfica de una ecuación cúbica.

un rayo de una rueda), ya en la prolongación de un radio (como en el reborde de la rueda de un tren). La figura 54 muestra estas dos curvas.

Se obtiene otra variante de la cicloide haciendo rodar la circunferencia, no sobre una recta, sino sobre otra circunferencia. Si el círculo rodante c, de radio r, es tangente interiormente al círculo mayor C,

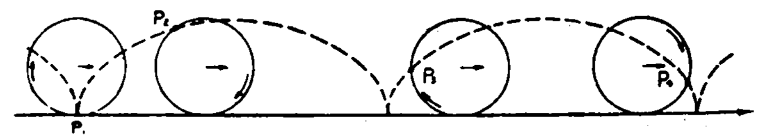


FIG. 53.—La cicloide.

de radio R, el lugar engendrado por un punto fijo de la circunferencia c se llama *hipocicloide*.

Si el círculo c describe la circunferencia entera de C una vez, el punto P retornará a su posición original sólo si el radio de C es un

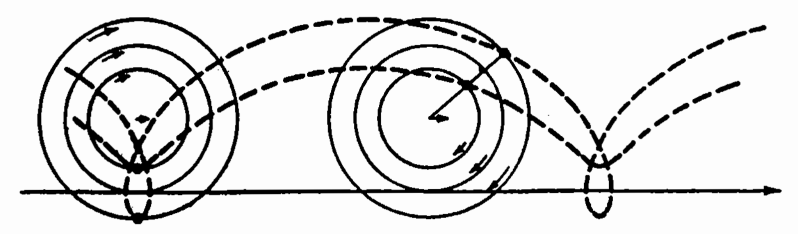


Fig. 54.—Cicloides generales.

múltiplo entero del de c. La figura 55 muestra el caso en que $R = 3r$. Con mayor generalidad, si el radio de C es m/n veces el de c, la hipocicloide se cerrará al cabo de n circuitos alrededor de C, y constará de m arcos. Un caso especial interesante se presenta si $R = 2r$. Todo

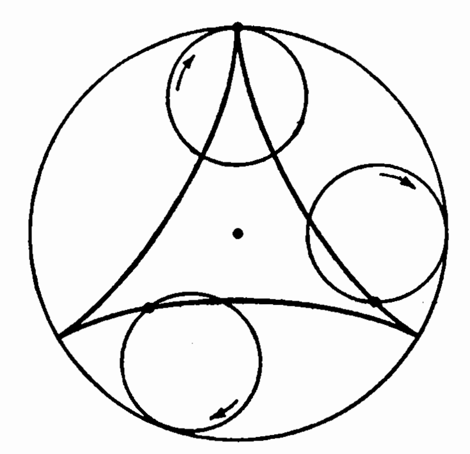


Fig. 55.—Hipocicloide de tres retrocesos.

punto P del círculo interior describe un diámetro del círculo mayor (Fig. 56). Proponemos como problema al lector la demostración de esta propiedad.

Aún puede engendrarse otro tipo de cicloide por medio de un círculo que rueda sobre otro fijo, permaneciendo tangente exteriormente a éste. Tal curva se llama *epicicloide*.

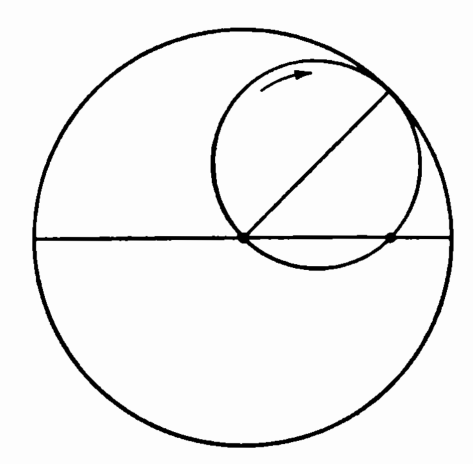


FIG. 56.—Movimiento rectilíneo por medio de puntos de un círculo que rueda sobre un círculo de radio doble.

***4. Conexiones. Inversores de Peaucellier y de Hart.**—Dejaremos por ahora el tema de las cicloides (que aparecerán nuevamente en un lugar inesperado) y estudiaremos otros métodos de engendrar curvas. Los instrumentos mecánicos más sencillos para trazar curvas son las *conexiones.* Una conexión consiste en un conjunto de varillas rígidas unidas de alguna manera mediante articulaciones móviles, de forma que el sistema total tenga suficiente libertad como para permitir a un punto del mismo describir una cierta curva. El compás es realmente una simple conexión, que consiste en principio en una varilla única fijada por un punto.

Las conexiones se han utilizado desde hace mucho tiempo en la construcción de máquinas. Uno de los ejemplos históricos famosos, el «paralelogramo de Watt», fué ideado por James Watt para resolver el problema de unir el pistón de su máquina de vapor con un punto de un volante, de forma que la rotación del volante hiciese mover el pistón en línea recta. La solución de Watt es sólo aproximada, y pese a los esfuerzos de muchos matemáticos distinguidos, el problema de construir una conexión que haga mover un punto *precisamente* en línea recta continuaba sin resolver. En un tiempo, cuando las demostraciones de irresolubilidad de ciertos problemas atraían enormemente la atención, fué formulada la conjetura de que la construcción de tal conexión era imposible. Fué por ello una gran sorpresa cuando, en 1864, un oficial naval francés, Peaucellier, inventó una sencilla conexión que resolvía el problema. Con la introducción de lubricantes eficaces, el

problema técnico para las máquinas de vapor había perdido entonces su importancia.

El propósito de la conexión de Peaucellier es el de convertir un

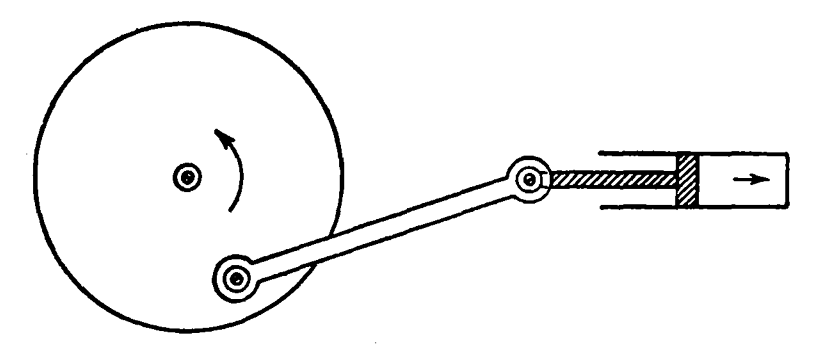


FIG. 57.—Movimiento rectilíneo transformado en rotación.

movimiento circular en rectilíneo. Está basado en la teoría de la inversión discutida en las páginas 153-57. Como se ve en la figura 58, la conexión consta de siete varillas rígidas; dos de longitud t, cuatro de longitud s, y la séptima de longitud arbitraria. O y R son dos pun-

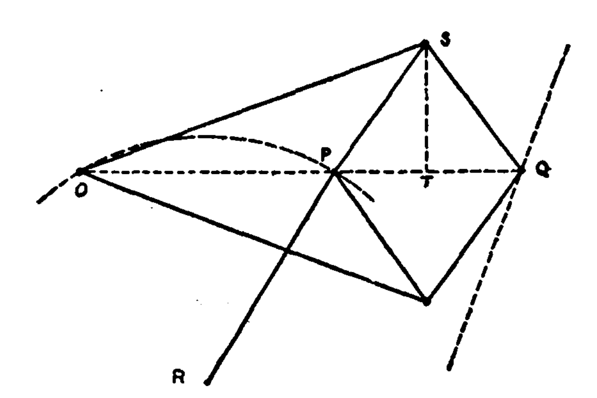


FIG. 58.—Transformación de Peaucellier de la rotación en movimiento rectilíneo.

tos fijos, situados de modo que $OR = PR$. El aparato entero puede moverse sujeto a las condiciones dadas. Vamos a demostrar que *si* P *describe una circunferencia de centro* R *con radio* PR, Q *describe un segmento de recta.* Designando el pie de la perpendicular de S a PQ por T, observemos que

$$OP \cdot OQ = (OT - PT)(OT + PT) = OT^2 - PT^2$$
$$= (OT^2 + ST^2) - (PT^2 + ST^2)$$
$$= t^2 - s^2.$$

La cantidad $t^2 - s^2$ es una constante que llamaremos r^2. Como $OP \cdot OQ = r^2$, P y Q son puntos inversos respecto a la circunferencia de radio r y centro O. Cuando P describe una trayectoria circular (que pasa por O), Q describe la curva inversa de la circunferencia, la cual es una recta, pues hemos visto que la inversa de una circunferencia que pasa por O es una recta. La trayectoria de Q es, por tanto, una recta, que se traza sin hacer uso de la regla. Otra conexión que resuelve el mismo problema es el inversor de Hart. Este consta de cinco varillas unidas como indica la figura 59. Aquí $AB=CD$, $BC=AD$. O, P y Q son puntos fijos de las varillas AB, AD y CB, respectivamente, tales que $AO/OB = AP/PD = CQ/QB = m/n$. Los puntos O y S son fijos, de forma que $OS = PS$, mientras que el resto del aparato puede moverse libremente. Es evidente que AC es constantemente paralela a BD; por tanto, O, P y Q son siempre colineales y OP es paralela a AC. Tracemos AE y CF perpendiculares a BD; tenemos:

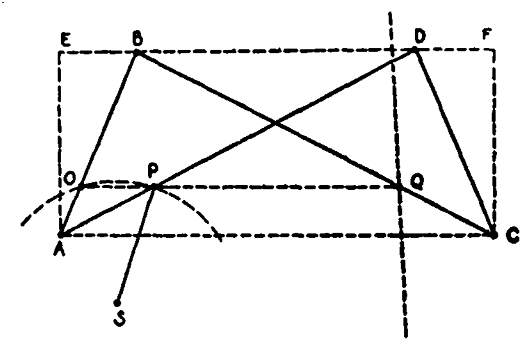


FIG. 59.—Inversor de Hart.

$$AC \cdot BD = EF \cdot BD = (ED + EB)(ED - EB) = ED^2 - EB^2.$$

Pero $ED^2 + AE^2 = AD^2$, y $EB^2 + AE^2 = AB^2$. Por tanto, $ED^2 - EB^2 = AD^2 - AB^2$. Ahora bien:

$$OP/BD = AO/AB = m/(m+n) \quad \text{y} \quad OQ/AC = OB/AB = n/(m+n).$$

En consecuencia,

$$OP \cdot OQ = [mn/(m+n)^2]\, BD \cdot AC = [mn/(m+n)^2]\,(AD^2 - AB^2).$$

Esta cantidad es la misma para todas las posiciones posibles de la conexión; por tanto, P y Q son puntos inversos respecto a cierta circunferencia de centro O. Cuando al aparato se mueve, P describe una circunferencia de centro S y que pasa por O, mientras su inverso Q describe una recta.

Pueden construirse otras conexiones (por lo menos en principio), para trazar elipses, hipérbolas, y, en general, toda curva dada por una ecuación algebraica $f(x, y) = 0$ de cualquier grado.

VI. COMPLEMENTOS SOBRE INVERSIÓN Y SUS APLICACIONES

1. Invariancia de ángulos. Haces de círculos.—Aunque la inversión respecto de una circunferencia cambia la forma de las figuras geométricas, es un hecho notable el que las nuevas figuras continúen poseyendo muchas de las propiedades de las figuras primitivas. Éstas son las propiedades que no cambian, o «invariantes» de la transformación. Como ya sabemos, la inversión transforma rectas y circunferencias en rectas y circunferencias; añadiremos ahora otra propiedad importante: *el ángulo entre dos rectas o curvas es invariante en la inversión.* Con esto queremos decir que dos curvas secantes se transforman por una inversión en otras dos curvas que se cortan bajo el mismo ángulo. Por ángulo entre dos curvas entendemos, naturalmente, el ángulo formado por sus tangentes.

La demostración puede estudiarse en la figura 60, que ilustra el caso especial de una curva C que corta a una recta OL en un punto P. La inversa C' de C corta a OL en el punto inverso P', el cual, como

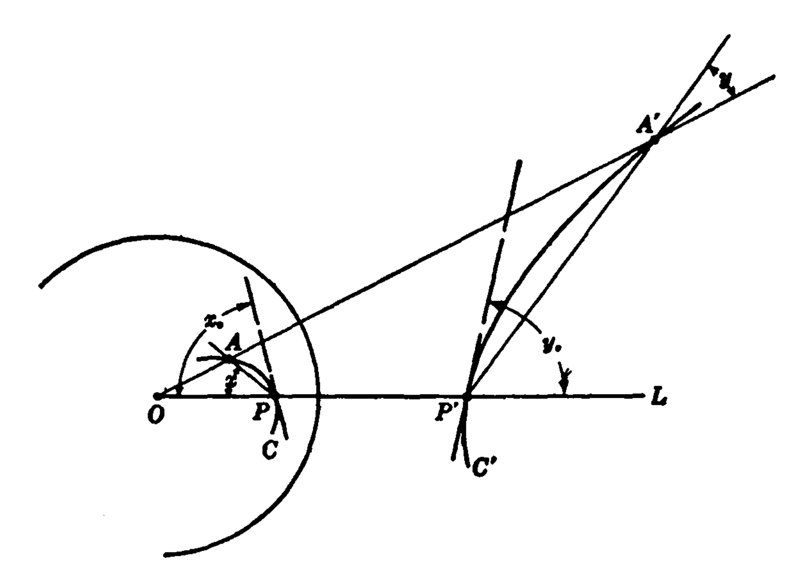


FIG. 60.—Conservación de los ángulos en la inversión.

OL es su propia inversa, está en OL. Demostraremos que el ángulo x_0 entre OL y la tangente a C en P es igual en magnitud al ángulo correspondiente y_0.

Para ello elegimos un punto A de la curva C próximo al P, y trazamos la secante AP. El inverso de A es un punto A' que, por estar a la vez en la recta OA y en la curva C', debe coincidir con su intersección. Dibujemos la secante $A'P'$. Por definición de inversión,

$$r^2 = OP \cdot OP' = OA \cdot OA',$$

o

$$\frac{OP}{OA} = \frac{OA'}{OP'};$$

es decir, los triángulos OAP y $OA'P'$ son semejantes. Por consiguiente, el ángulo x es igual al $OA'P'$, que llamaremos y. La etapa final consiste en hacer mover el punto A sobre C, aproximándolo al P; esto obliga a la secante AP a girar hasta la posición de la tangente a la curva C en P, mientras el ángulo x tiende al x_0. Al mismo tiempo A' se aproxima a P', y $A'P'$ gira hasta la posición de la tangente a C' en P'; el ángulo y tiende al y_0. Como x e y son iguales en cualquier posición de A, tendremos en el límite: $x_0 = y_0$.

Nuestra demostración queda incompleta, sin embargo, pues hemos considerado sólo el caso de una curva que corta a una recta que pasa por O. El caso general de dos curvas C y C^*, que forman un ángulo z en P, es ahora fácil de estudiar. Es evidente que la recta OPP' divide a z en dos ángulos, cada uno de los cuales se conserva en la inversión.

Debe observarse que aunque la inversión conserva la *magnitud* de los ángulos, invierte su *sentido*; es decir, si un rayo por P describe el ángulo x_0 en sentido contrario al de las agujas del reloj, su imagen recorrerá el ángulo y_0 en el sentido opuesto.

Una consecuencia particular de la invariancia de los ángulos en la inversión es que si dos circunferencias o curvas son *ortogonales* (es decir, se cortan en ángulo recto), continúan siendo ortogonales después de la inversión, mientras que dos curvas tangentes (es decir, que forman ángulo nulo) quedan tangentes.

Consideremos la familia de todas las circunferencias que pasan por el centro de inversión O y por otro punto fijo A del plano. Por lo dicho en la página 155, sabemos que esta familia de circunferencias se transformará en un haz de rectas que pasan por A'. Las circunferencias ortogonales a las primeras se transforman en circunferencias

ortogonales a las rectas del haz A', como se ve en la figura 61. (Las circunferencias ortogonales están dibujadas de trazos). El aspecto sencillo del haz de rectas resulta muy diferente del de la familia de cir-

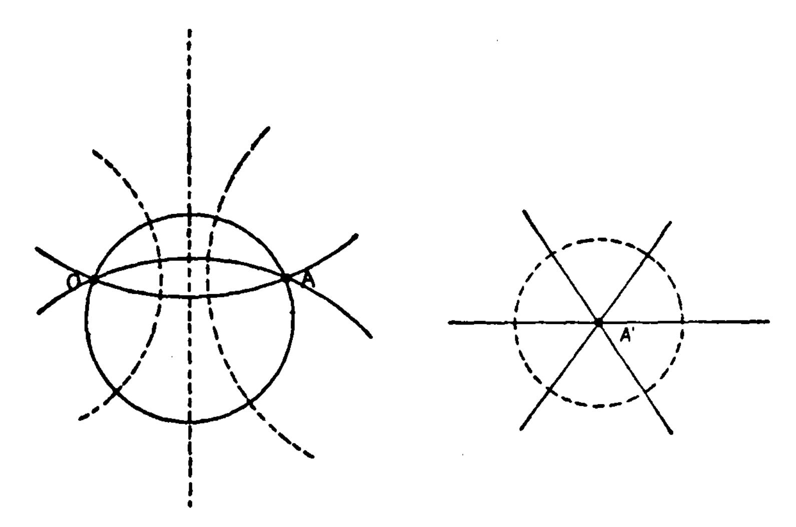


FIG. 61.—Haces de círculos ortogonales relacionados por inversión.

cunferencias, pese a que vemos que están íntimamente relacionados, pues, en efecto, desde el punto de vista de la inversión, son por completo equivalentes.

Otro ejemplo del efecto de la inversión lo constituye una familia de círculos tangentes entre sí en el punto de inversión. Efectuada la

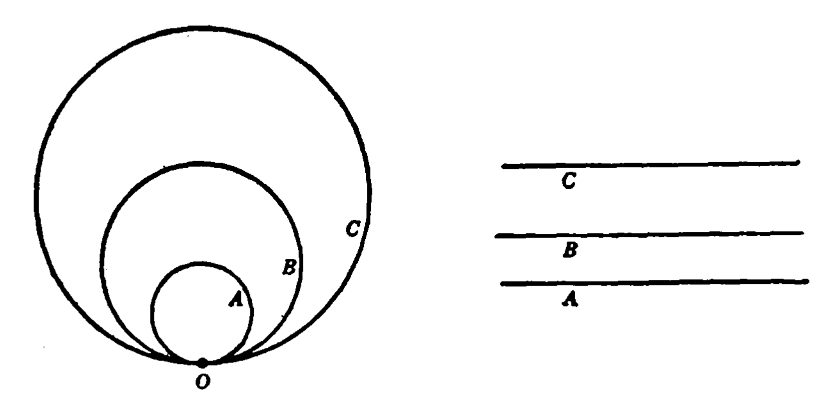


FIG. 62.—Circunferencias tangentes transformadas en rectas paralelas.

transformación se tendrá un haz de rectas paralelas, pues las imágenes de las circunferencias son rectas y ningún par de éstas se corta, ya que las circunferencias originales no tienen otro punto común que el O.

2. Aplicación al problema de Apolonio.—Una buena ilustración de la utilidad de la inversión es la siguiente solución geométrica inmediata del problema de Apolonio. Mediante una inversión respecto a un punto cualquiera, el problema de Apolonio para tres círculos dados puede transformarse en el problema de Apolonio correspondiente a otros tres círculos (¿por qué?). Luego, si podemos resolver el problema para una terna cualquiera de circunferencias, quedará resuelto para otra terna cualquiera, obtenida de la primera por inversión. Sa-

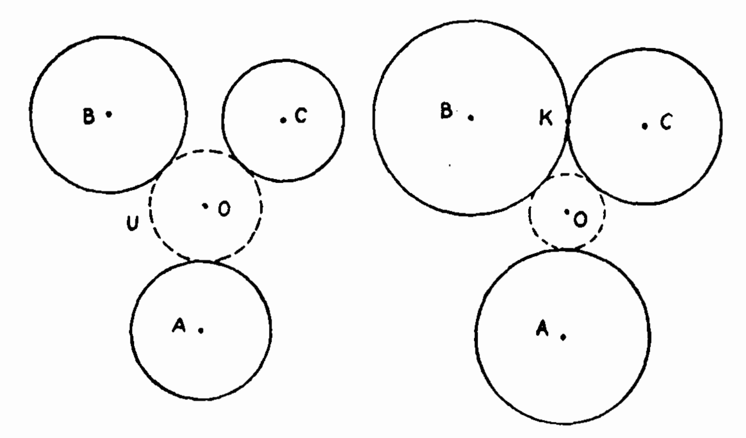


FIG. 63.—Construcción preliminar de Apolonio.

caremos ventaja de este hecho eligiendo entre todas estas ternas de círculos equivalentes una determinada, para la cual el problema resulta casi trivial.

Partimos de tres circunferencias de centros A, B, C, y supondremos que la circunferencia buscada U, de centro O y radio ρ, es tangente exteriormente a los tres círculos dados. Si aumentamos los radios de los tres círculos dados en una misma cantidad d, entonces la circunferencia de igual centro O y radio $\rho - d$ resolverá, evidentemente, el nuevo problema. Como fase inicial haremos uso de este hecho para reemplazar las tres circunferencias dadas por otras tres, tales que dos de ellas sean tangentes entre sí en un punto K (Fig. 63). A continuación, invertiremos la figura entera respecto de un círculo de centro K. Las circunferencias de centros B y C se transformarán en rectas paralelas b y c, mientras la tercera se transformará en otra

circunferencia *a* (Fig. 64). Sabemos que *a*, *b* y *c* pueden ser construídos con regla y compás. La circunferencia desconocida *U* se transformará en una circunferencia *u* tangente a *a*, *b* y *c*, y su radio, *r*, es evi-

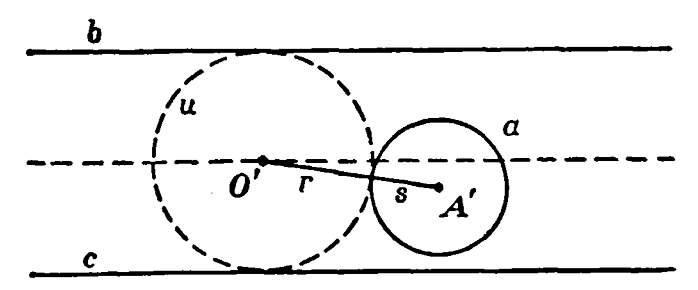


FIG. 64.—Solución del problema de Apolonio.

dentemente la mitad de la distancia entre *b* y *c*. Su centro *O'* es una de las dos intersecciones de la paralela media entre *b* y *c* con la circunferencia de centro *A'* (centro de *a*) y radio $r + s$ (*s* es el radio de *a*). Finalmente, por construcción de la circunferencia inversa de *u*, determinaremos el centro del círculo de Apolonio buscado, *U*. (Su centro *O* será el inverso, en la circunferencia de inversión, del punto inverso de *K* en *u*.)

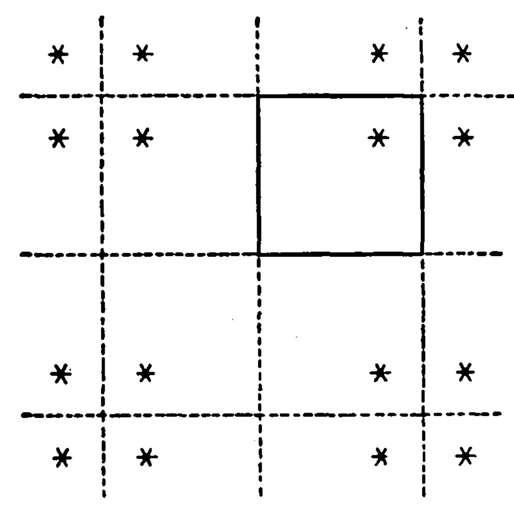


FIG. 65.—Simetría reiterada sobre los cuatro lados de un cuadrado.

***3. Simetrías reiteradas.**—Para todos son familiares los extraños fenómenos de reflexión que se producen cuando se dispone de más de un espejo. Si las cuatro paredes de una habitación rectangular estuvieran cubiertas con cuatro espejos ideales no absorbentes, un punto luminoso tendría infinitas imágenes, cada una correspondiente a cada habitación congruente obtenida por reflexión (figura 65). Un conjunto de espejos menos regular (p. ej., tres espejos) dará

una serie mucho más complicada de imágenes. La configuración resultante puede ser descrita fácilmente sólo si los triángulos reflejados forman un cubrimiento sin solapamiento del plano. Esto sucede sólo

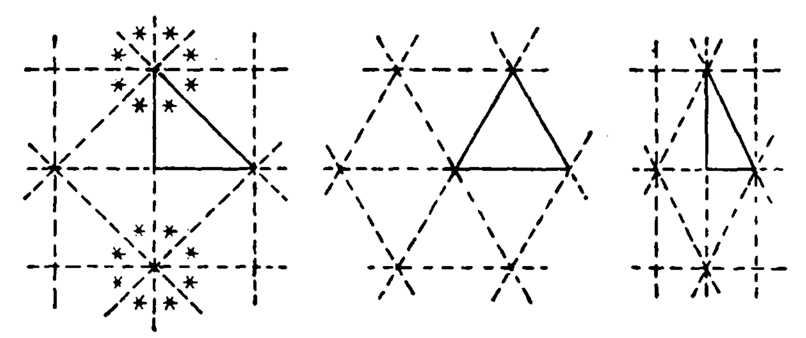


FIG. 66.—Constelaciones regulares de espejos triangulares.

en los casos del triángulo isósceles rectángulo, el triángulo equilátero y la mitad rectangular de éste. (Véase Fig. 66.)

La situación se hace mucho más interesante si consideramos inversiones repetidas respecto a un par de circunferencias. Colocándonos entre dos espejos circulares concéntricos se pueden ver otros infinitos

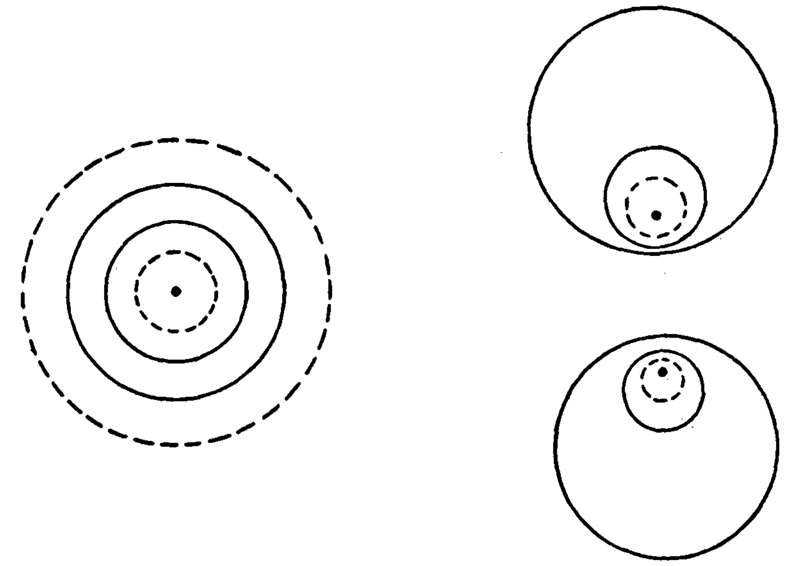


FIG. 67.—Reflexión reiterada en sistemas de dos círculos.

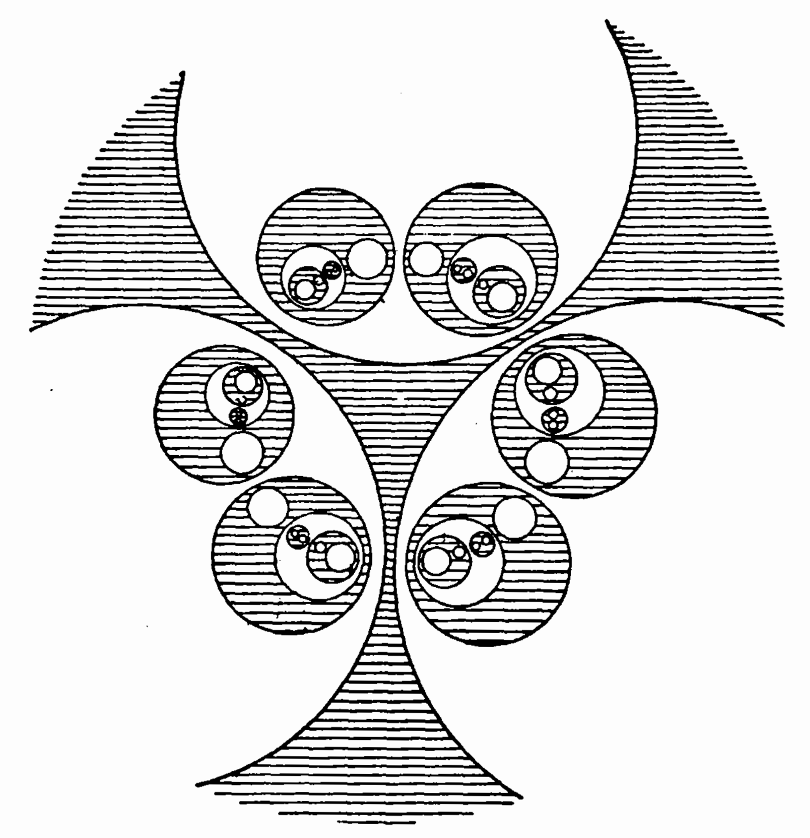


FIG. 68.—Reflexión en un sistema de tres círculos.

círculos concéntricos. Una sucesión de estos círculos tiende a infinito, mientras la otra se concentra alrededor del centro.

El caso de dos circunferencias exteriores es algo más complicado. Aquí los círculos y sus imágenes se reflejan sucesivamente uno en otro disminuyendo en cada reflexión, hasta reducirse a dos puntos, uno en cada círculo. (Estos puntos tienen la propiedad de ser inversos respecto a ambos círculos; véase figura 67.) Con tres círculos se forma el bello dibujo de la figura 68.

CAPÍTULO IV

GEOMETRÍA PROYECTIVA. AXIOMÁTICA. GEOMETRÍAS NO EUCLÍDEAS

I. INTRODUCCIÓN

1. Clasificación de las propiedades geométricas. Invariancia respecto a las transformaciones.—La geometría se ocupa de las propiedades de las figuras del plano o del espacio. Estas propiedades son tan numerosas y variadas, que es necesario algún principio de clasificación para poner orden en esta riqueza de conocimientos. Se puede, p. ej., introducir una clasificación basada en el método utilizado para deducir los teoremas. Desde este punto de vista, se hace usualmente una distinción entre procedimientos «sintéticos» y «analíticos». El primero de éstos es el método axiomático clásico de Euclides, en el cual el edificio se construye sobre fundamentos puramente geométricos, independientes del álgebra y del concepto de continuo numérico, y los teoremas se deducen por razonamiento lógico de un conjunto inicial de proposiciones llamadas axiomas o postulados. El segundo método se basa en la introducción de coordenadas numéricas, y utiliza la técnica del álgebra. Este método ha originado un cambio profundo en la ciencia matemática, del que ha resultado la unificación de la geometría, el análisis y el álgebra en un sistema orgánico.

En este capítulo, la clasificación según el método será menos importante que la hecha de acuerdo con el *contenido*, basada en el carácter de los propios teoremas, e independiente de los métodos usados para demostrarlos. En geometría plana elemental se distingue entre teoremas que tratan de la congruencia de figuras y utilizan los conceptos de longitud y ángulo, y teoremas que se refieren a la semejanza de figuras, y que sólo utilizan el concepto de ángulo. Esta distinción particular no es muy importante, ya que longitud y ángulo están relacionados tan íntimamente que más bien resulta artificial separarlos. (Es el estudio de esta relación lo que constituye la mayor parte del contenido de la trigonometría.) En lugar de ello, podemos decir que los teoremas de la geometría elemental se ocupan de *magnitudes*: longitudes, medidas de ángulos, y áreas. Dos figuras son equivalentes, desde este punto de vista, si son *congruentes*, es decir, si una puede obtenerse de la otra mediante un *movimiento rígido*, en el cual sólo

cambia la posición, pero no la magnitud. Surge ahora la cuestión de si el concepto de magnitud y los de congruencia y semejanza, relacionados con él, son esenciales en geometría, o si las figuras geométricas pueden tener aún otras propiedades más profundas, que subsistan después de transformaciones más drásticas que los movimientos rígidos. Vamos a ver que así acontece.

Supongamos que se dibuja una circunferencia y un par de diámetros perpendiculares en un trozo rectangular de madera blanda, como indica la figura 69. Si colocamos el trozo entre las mordazas de un potente tornillo de banco y lo comprimimos hasta que tenga la mitad de su ancho original, el círculo se habrá transformado en una

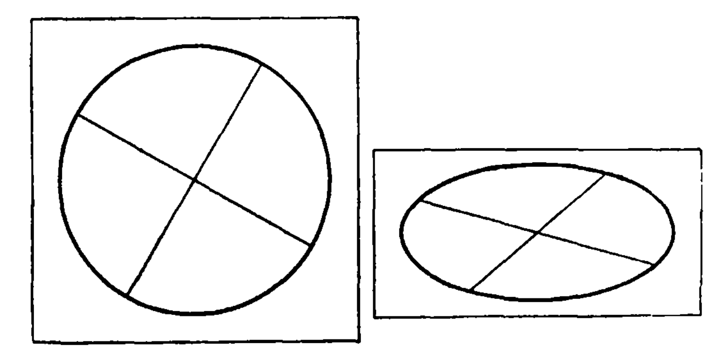


FIG. 69.—Compresión de un círculo.

elipse y el ángulo entre los diámetros de la elipse ya no será recto. La circunferencia tiene la propiedad de que sus puntos equidistan del centro, mientras que esto no es verdad para la elipse. Pudiera parecer que todas las propiedades geométricas de la figura original han sido destruídas por la compresión. Pero ni mucho menos es así; p. ej., la propiedad de que el centro divide a cada diámetro en dos partes iguales es válida, tanto para la circunferencia como para la elipse. He aquí una propiedad que subsiste después del cambio drástico en las magnitudes de la figura original. Esta observación sugiere la posibilidad de clasificar los teoremas sobre las figuras geométricas según que sigan verificándose o dejen de ser ciertos cuando la figura se somete a una compresión uniforme. Más en general: dado un tipo definido de transformaciones de una figura (como la clase de los movimientos rígidos, compresiones, inversiones respecto a circunferencias, etc.) podemos investigar qué propiedades de la figura permanecen invariables cuando se somete a esta clase de transformaciones. El conjunto de teore-

mas referentes a estas propiedades será la *geometría asociada a dicha clase de transformaciones.* La idea de clasificar las diferentes ramas de la geometría de acuerdo con los tipos de transformaciones consideradas fué propuesta por Félix Klein (1849-1925) en su famosa comunicación (el «Programa de Erlangen») presentada en 1872. Desde entonces ha dominado enormemente el pensamiento matemático.

En el capítulo V nos encontraremos con el hecho verdaderamente sorprendente de que ciertas propiedades geométricas son tan íntimamente intrínsecas, que persisten después que las figuras han sido sometidas a deformaciones muy arbitrarias; figuras dibujadas en una lámina de caucho que se estira o comprime de cualquier manera, todavía conservan algunas de sus características originales. En este capítulo, sin embargo, vamos a ocuparnos solamente de aquellas propiedades que permanecen invariables o «invariantes» bajo un tipo especial de transformaciones, situado entre la clase muy restringida de los movimientos rígidos, por un lado, y la clase más general de las deformaciones arbitrarias, por otro. Ésta es la clase de las «transformaciones proyectivas».

2. Transformaciones proyectivas.—Los matemáticos se vieron impulsados desde hace mucho tiempo al estudio de estas propiedades geométricas, debido a los problemas de *perspectiva,* que fueron estudiados por artistas como Leonardo de Vinci y Alberto Durero. La imagen trazada por un pintor puede considerarse como la proyección del original sobre la tela, con el centro de proyección en el ojo del pintor. En este proceso, las longitudes y los ángulos se alteran necesariamente en forma que depende de las posiciones relativas de los diversos objetos pintados. Sin embargo, puede reconocerse sobre la tela, generalmente, la estructura geométrica del original. ¿Cómo es esto posible? Lo es, porque existen propiedades geométricas «invariantes en la proyección», propiedades que aparecen sin alterar en la imagen y que hacen posible la identificación. Deducir y analizar estas propiedades constituye el objeto de la geometría proyectiva.

Es obvio que los teoremas de esta rama de la geometría no pueden ser proposiciones sobre longitudes y ángulos o sobre congruencia. Algunos hechos aislados, de naturaleza proyectiva, son bien conocidos desde el siglo XVII, y aun, como en el caso del «teorema de Menelao», desde la antigüedad clásica. Pero el estudio sistemático de la geometría proyectiva no comenzó hasta fines del siglo XVIII, cuando l'École Polytechnique de París inició una nueva etapa de progreso matemático, particularmente en geometría. Esta Escuela, surgida de la Revolución francesa, produjo muchos oficiales para los servicios militares de la

República. Uno de sus graduados fué J. V. Poncelet (1788-1867); que escribió su famoso *Traité des propriétés projectives des figures* en 1813, mientras era prisionero de guerra en Rusia. En el siglo XIX, bajo la influencia de Steiner, von Staudt, Chasles y otros, la geometría proyectiva se convirtió en uno de los principales temas de investigación matemática. Su popularidad fué debida en parte a su gran encanto estético y en parte también a su efecto aclaratorio sobre la geometría en conjunto, y a su íntima conexión con la geometría no euclídea y el álgebra.

II. CONCEPTOS FUNDAMENTALES

1. Grupo de las transformaciones proyectivas.—Definiremos primero la clase o *grupo* [1] de las transformaciones proyectivas. Supongamos dos planos π y π' en el espacio, no necesariamente paralelos, y

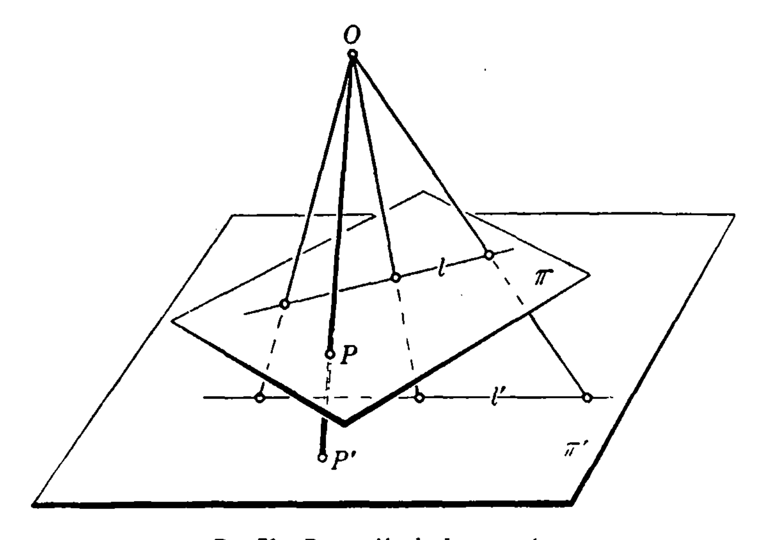


FIG. 70.—**Proyección desde un punto.**

hagamos una *proyección central* de π sobre π' desde un centro dado O no situado sobre π ni sobre π', definiendo la imagen de cada punto P de π como aquel punto P' de π' tal que P y P' están sobre la misma

[1] La palabra *grupo*, cuando se aplica a una clase de transformaciones, implica que la aplicación sucesiva de dos transformaciones de la clase equivale a una transformación de la misma clase, y que la *inversa* de una transformación de la clase pertenece también a ella. Las propiedades de grupo de las operaciones matemáticas han desempeñado y seguirán desempeñando un gran papel en muchos campos, aunque en geometría, quizá, la importancia del concepto de grupo ha sido algo exagerada.

recta que pasa por O. Podemos también efectuar una *proyección paralela* si hacemos que todos los rayos proyectantes sean paralelos. Del mismo modo se puede definir la proyección de una recta l de un plano π sobre otra recta l' de π' desde un punto O de π, o hacerlo mediante proyección paralela.

Una representación de una figura sobre otra mediante una proyección central o paralela, o por una sucesión finita de tales proyecciones, se llama *transformación proyectiva*[1]. La *geometría proyectiva* del plano

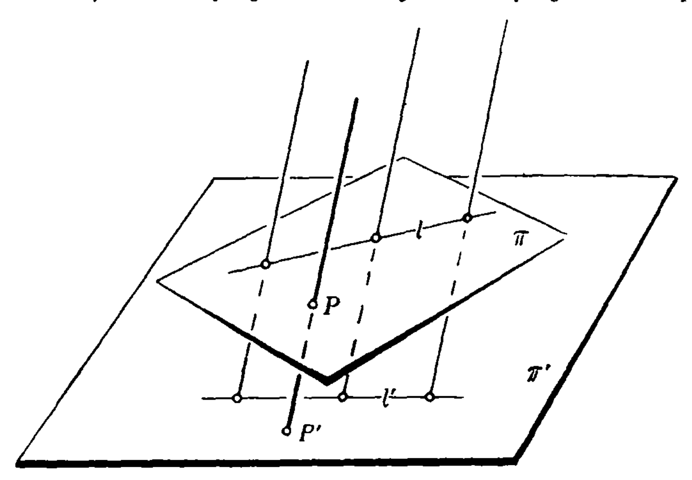


FIG. 71.—Proyección paralela.

o de la recta consiste en el conjunto de aquellas proposiciones geométricas que permanecen invariables en una transformación proyectiva arbitraria de las figuras a las que se refieren. Por el contrario, llamaremos *geometría métrica* al conjunto de aquellas proposiciones que tratan de las magnitudes de las figuras, invariantes sólo respecto a la clase de los movimientos rígidos.

Algunas propiedades proyectivas pueden reconocerse inmediatamente. Un punto, por supuesto, se proyectará en un punto. Además, *una recta se proyectará en una recta*; pues si la recta l de π se proyecta sobre el plano π', la intersección de π' con el plano determinado por O y l será una recta [2]. Si un punto A y una recta l son incidentes [3], el

[1] Dos figuras relacionadas por una *sola* proyección se dicen comúnmente *perspectivas*. Así, pues, una figura F está ligada mediante una transformación proyectiva a otra F' si F y F' son perspectivas, o si podemos hallar una sucesión de figuras $F, F_1, F_2, F_3, \ldots F_n, F'$, tales que cada una sea perspectiva con la siguiente.

[2] Hay excepciones si la recta OP es paralela al plano π (o si lo es el plano de O y l). Estas excepciones serán tratadas en IV.

[3] Un punto y una recta se dicen incidentes si la recta pasa por el punto o si el punto está en la recta.

punto correspondiente A' y la recta l' lo son también en la proyección. Luego la *incidencia de punto y recta es invariante respecto al grupo proyectivo.* De este hecho surgen muchas consecuencias simples, pero importantes. Si tres o más puntos son *colineales*, es decir, pertenecen a una misma recta, entonces sus imágenes son también colineales. Asimismo, si en el plano π tres o más rectas son *concurrentes*, es decir, inciden en un mismo punto, entonces sus imágenes serán también rectas concurrentes. Mientras estas propiedades simples, incidencia, colinealidad y concurrencia son *propiedades proyectivas* (es decir, propie-

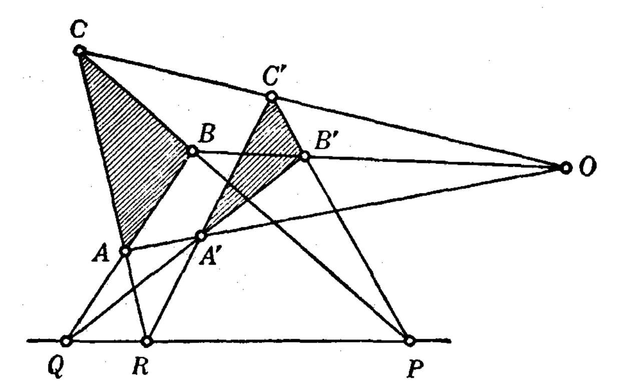


FIG. 72.—Configuración de Desargues en el plano.

dades invariantes en las proyecciones), en cambio, las medidas de longitudes y ángulos, y las razones de tales magnitudes, se alteran en general en la proyección. Los triángulos equiláteros e isósceles pueden proyectarse en triángulos cuyos lados tengan longitudes diferentes. Por tanto, aunque «triángulo» es un concepto de geometría proyectiva, «triángulo equilátero» no lo es, y pertenece sólo a la geometría métrica.

2. Teorema de Desargues.—Uno de los primeros descubrimientos de la geometría proyectiva fué el famoso teorema sobre triángulos de Desargues (1593-1662). *Si, en un plano, dos triángulos* ABC *y* A'B'C' *son tales que las rectas que unen vértices correspondientes concurren en un punto* O, *los lados correspondientes se cortan en tres puntos colineales.* La figura 72 ilustra el teorema y el lector puede dibujar otras figuras para comprobarlo experimentalmente. La demostración no es trivial, no obstante la sencillez de la figura, constituída sólo por rectas. El teorema pertenece sin duda a la geometría proyectiva, pues si proyectamos la figura entera sobre otro plano, conservará todas las propiedades enunciadas en el teorema. Daremos una demostración de

este teorema en V. Por el momento, observemos el hecho notable de que el teorema de Desargues es también cierto si los triángulos están en *diferentes* planos (no paralelos), y que este teorema de Desargues de la geometría tridimensional es muy fácil de probar.

Supongamos que las rectas AA', BB' y CC' se cortan en O (figura 73), de acuerdo con la hipótesis. Entonces AB está en el mismo plano de $A'B'$, de modo que estas dos rectas se cortan en un punto Q; análogamente AC y $A'C'$ se cortan en R, y BC y $B'C'$ se cortan en P. Como P, Q y R están en las prolongaciones de los lados de ABC y $A'B'C'$, deben estar en el mismo plano de cada uno de los dos triángulos, y, en consecuencia, están sobre la recta de intersección de estos

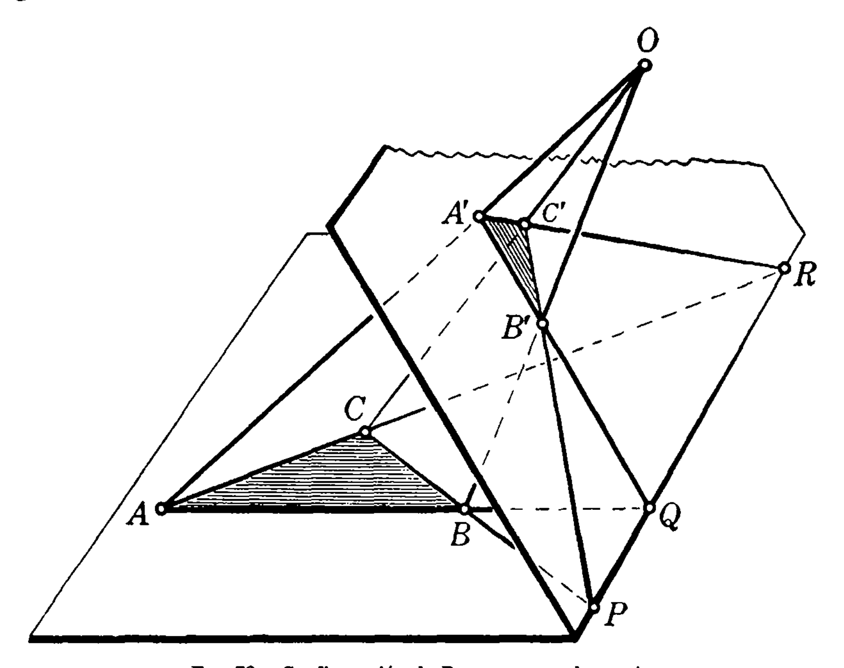


FIG. 73.—Configuración de Desargues en el espacio.

dos planos. Por tanto, P, Q y R son colineales, como queríamos probar.

Esta sencilla demostración sugiere que podemos probar el teorema para dos dimensiones mediante, por decirlo así, un paso al límite, aplastando la figura total de manera que los dos planos coincidan en el límite y el punto O, junto con los otros, caiga sobre este plano. Sin embargo, hay cierta dificultad en llevar a cabo tal proceso de límite, pues la recta de intersección PQR no está unívocamente determinada

cuando los planos coinciden. No obstante, la figura 72 puede considerarse como una perspectiva de la figura del espacio representada en la 73, y este hecho puede utilizarse para demostrar el teorema en el caso plano.

En efecto, existe una diferencia fundamental entre el teorema de Desargues en el plano y en el espacio. Nuestra demostración en tres dimensiones utiliza razonamientos geométricos basados solamente en los conceptos de incidencia e intersección de puntos, rectas y planos. Puede probarse que la demostración del teorema bidimensional, *si se procede por completo en el plano*, requiere necesariamente el uso del concepto de semejanza de figuras, que es equivalente al concepto métrico de longitud y no es ya una noción proyectiva.

El teorema *recíproco* del de Desargues establece que si ABC y $A'B'C'$ son dos triángulos situados de modo que los puntos de intersección de los pares de lados correspondientes sean colineales, las rectas que unen vértices correspondientes son concurrentes. Su demostración para el caso en que los triángulos estén en dos planos no paralelos, se deja como ejercicio al lector.

III. RAZÓN DOBLE

1. Definición y prueba de su invariancia.—Así como la longitud de un segmento rectilíneo es la clave de la geometría métrica, hay también un concepto fundamental de la geometría proyectiva, mediante el cual pueden expresarse todas las propiedades netamente proyectivas de las figuras.

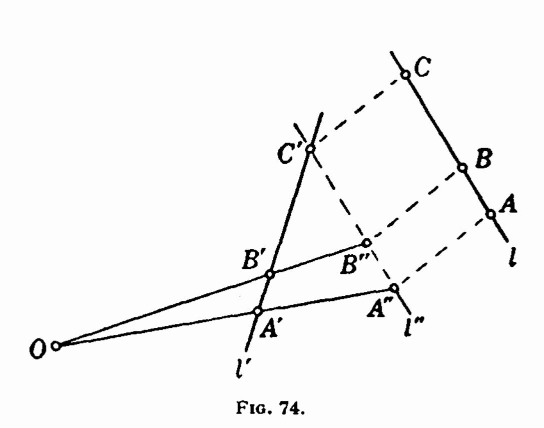


FIG. 74.

Si tres puntos A, B, C están sobre una recta, la proyección cambia, en general, no sólo las distancias AB y BC, sino también la razón AB/BC. En efecto, tres puntos cualesquiera A, B, C de una recta l pueden ser siempre coordinados con otros tres puntos arbitrarios A', B', C' de otra recta l' por medio de dos proyecciones sucesivas. Para esto, hagamos girar la recta l' alrededor del punto C', hasta que tome la posición l'' paralela a l (véase Fig. 74). A continuación, proyectamos l sobre l'', mediante una proyección paralela a la recta que une C con C', definiendo tres puntos A'', B''

y C'' ($C'' = C'$). Las rectas $A'A''$ y $B'B''$ deben cortarse en un punto O que elegimos como centro de la segunda proyección. Estas dos proyecciones nos dan el resultado deseado[1].

Según acabamos de ver, ninguna magnitud relativa sólo a tres puntos de una recta queda invariante en una proyección. Pero —y esto constituye el descubrimiento decisivo de la geometría proyectiva— si tenemos *cuatro* puntos A, B, C, D de una recta, y los proyectamos sobre otra recta en A', B', C', D', existe entonces una cierta magnitud, llamada *razón doble* de los cuatro puntos, que conserva su valor en la proyección. He aquí una propiedad matemática de un conjunto de cuatro puntos de una recta que no desaparece mediante proyección y que puede reconocerse en cualquier imagen de la recta. La razón doble no es ni una longitud ni un cociente de dos longitudes, sino *la razón de dos cocientes*; si consideramos los cocientes CA/CB y DA/DB, por definición, la razón doble de los cuatro puntos A, B, C, D, tomados en este orden, es:

$$x = \frac{CA}{CB} : \frac{DA}{DB}$$

A B C D

Vamos a demostrar que *la razón doble de cuatro puntos es invariante en la proyección*; es decir, que si A, B, C, D y A', B', C', D' son puntos correspondientes de dos rectas relacionadas por proyección, se verifica

$$\frac{CA}{CB} : \frac{DA}{DB} = \frac{C'A'}{C'B'} : \frac{D'A'}{D'B'}$$

La demostración se logra con medios elementales. Recordemos que el área de un triángulo es igual a $^1/_2$ (base $\times$ altura) y está también dada por la mitad del producto de dos cualesquiera de sus lados por el seno del ángulo comprendido. Tenemos, en la figura 75:

$$\text{área } OCA = {}^1/_2 h \cdot CA = {}^1/_2 OA \cdot OC \text{ sen } \widehat{COA}$$

$$\text{área } OCB = {}^1/_2 h \cdot CB = {}^1/_2 OB \cdot OC \text{ sen } \widehat{COB}$$

$$\text{área } ODA = {}^1/_2 h \cdot DA = {}^1/_2 OA \cdot OD \text{ sen } \widehat{DOA}$$

$$\text{área } ODB = {}^1/_2 h \cdot DB = {}^1/_2 OB \cdot OD \text{ sen } \widehat{DOB},$$

[1] ¿Qué sucede si las rectas $A'A''$ y $B'B''$ son paralelas?

de donde se deduce:

$$\frac{CA}{CB}:\frac{DA}{DB}=\frac{CA}{CB}\cdot\frac{DB}{DA}=\frac{OA\cdot OC\cdot \text{sen}\,\widehat{COA}}{OB\cdot OC\cdot \text{sen}\,\widehat{COB}}\cdot\frac{OB\cdot OD\cdot \text{sen}\,\widehat{DOB}}{OA\cdot OD\cdot \text{sen}\,\widehat{DOA}}$$

$$=\frac{\text{sen}\,\widehat{COA}}{\text{sen}\,\widehat{COB}}\cdot\frac{\text{sen}\,\widehat{DOB}}{\text{sen}\,\widehat{DOA}}$$

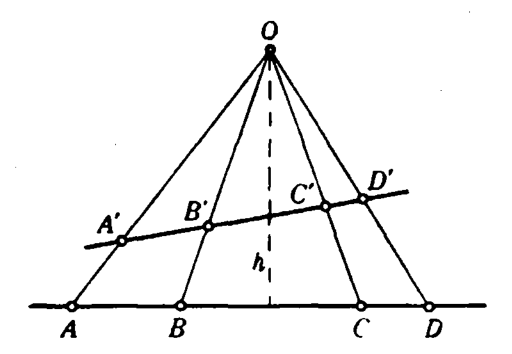


Fig. 75.—Conservación de la razón doble en la proyección central.

Luego la razón doble de A, B, C, D depende únicamente de los ángulos subtendidos desde O por los segmentos que unen A, B, C, D. Como estos ángulos son los mismos para otros cuatro puntos cualesquiera A', B', C', D' en los que A, B, C, D puedan ser proyectados desde O, resulta que la razón doble permanece invariable en la proyección.

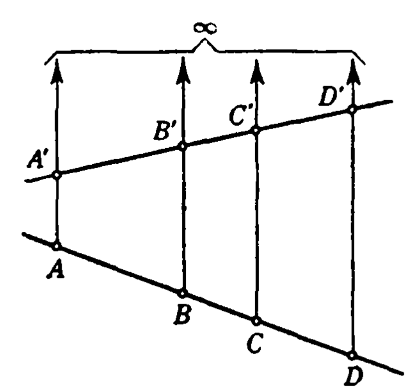


Fig. 76.—Conservación de la razón doble en la proyección paralela.

Que la razón doble de cuatro puntos no varía en una proyección *paralela* se obtiene de propiedades elementales de los triángulos semejantes, y la demostración se deja como ejercicio al lector.

Hasta aquí, hemos entendido que la razón doble de cuatro puntos A, B, C, D de una recta l sólo incluye longitudes positivas. Resulta conveniente modificar esta definición como sigue: elegimos un sentido sobre l como positivo, y convenimos en que las longitudes

medidas en este sentido serán positivas, mientras que las medidas en sentido opuesto serán negativas. Definiremos entonces la razón doble de A, B, C, D, considerados en este orden, por

$$(ABCD) = \frac{CA}{CB} : \frac{DA}{DB}, \qquad [1]$$

donde los números CA, CB, DA, DB han de ser tomados con su propio signo. Como la inversión del sentido positivo de l sólo cambia el signo de cada término de este cociente, el valor de $(ABCD)$ no dependerá del sentido elegido como positivo. Es fácil ver que $(ABCD)$ será negativa o positiva según que el par de puntos A, B esté o no separado (es decir, entrelazado) por el par C, D. Como esta propiedad de separación es invariante en la proyección, el signo de la razón doble $(ABCD)$ es también invariante. Si elegimos un punto fijo O de l como origen, y tomamos como abscisa x de cada punto de l su distancia orientada a partir de O; es decir, si las abscisas de A, B, C, D son x_1, x_2, x_3, x_4, respectivamente, tenemos que

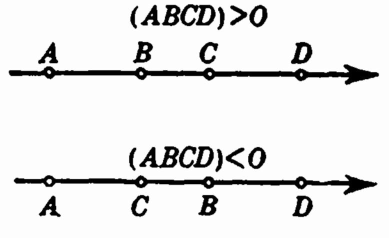


FIG. 77.—Signo de la razón doble.

$$(ABCD) = \frac{CA}{CB} : \frac{DA}{DB} = \frac{x_3 - x_1}{x_3 - x_2} : \frac{x_4 - x_1}{x_4 - x_2} = \frac{x_3 - x_1}{x_3 - x_2} \cdot \frac{x_4 - x_2}{x_4 - x_1}$$

Si $(ABCD) = -1$, es decir, $CA/CB = -DA/DB$, C y D dividen entonces al segmento AB, interior y exteriormente, en la misma razón. En este caso se dice que C y D dividen al segmento AB *armónicamente*, y cada uno de los puntos C y D se llama *conjugado armónico* del otro, respecto del par A, B. Si $(ABCD) = 1$, los puntos C y D (o A y B) coinciden.

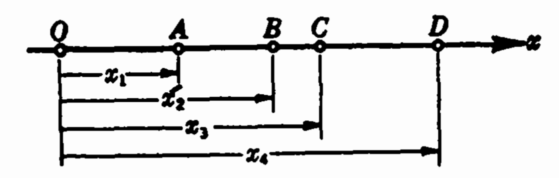


FIG. 78.—Razón doble expresada por medio de abscisas.

Se debe tener presente que el orden en que se toman A, B, C, D

es parte esencial de la definición de la razón doble $(ACBD)$; p. ej., si $(ABCD) = \lambda$, la razón doble $(BACD)$ es $1/\lambda$, mientras que $(ACBD) = 1 - \lambda$, como el lector puede verificar fácilmente. Cuatro puntos A, B, C, D pueden ordenarse de $4 \cdot 3 \cdot 2 \cdot 1 = 24$ maneras distintas, cada una de las cuales da un cierto valor a la razón doble. Algunas de estas permutaciones dan, no obstante, el mismo valor que el orden original A, B, C, D; por ejemplo, $(ABCD) = (BADC)$. Dejamos como ejercicio al lector el demostrar que hay sólo seis valores diferentes de la razón doble para las 24 permutaciones diferentes de los puntos, que son:

$$\lambda, \quad 1-\lambda, \quad 1/\lambda, \quad \frac{\lambda-1}{\lambda}, \quad \frac{1}{1-\lambda}, \quad \frac{\lambda}{\lambda-1}$$

Estas seis cantidades son, en general, distintas, pero dos de ellas pueden coincidir, como en el caso de la división armónica, cuando $\lambda = -1$.

Podemos definir también la *razón doble de cuatro rectas* 1, 2, 3, 4 *coplanarias* (es decir, situadas en un mismo plano) y *concurrentes*, como la razón doble de los cuatro puntos de intersección de estas rectas con otra del mismo plano. La posición de esta quinta recta es indiferente, debido a la invariancia de la razón doble en la proyección. Equivalente a ésta es la definición:

$$(1\,2\,3\,4) = \frac{\text{sen}\,(1,3)}{\text{sen}\,(2,3)} : \frac{\text{sen}\,(1,4)}{\text{sen}\,(2,4)},$$

que debe tomarse con signo más o menos según que un par de rectas separe o no al otro par. [En esta fórmula (1, 3), p. ej., representa el ángulo formado por las rectas 1 y 3.] Finalmente, podemos definir la *razón doble de cuatro planos coaxiales* (cuatro planos del espacio que se cortan en una recta l, su eje). Si una recta corta a los planos en cuatro puntos, estos puntos tendrán siempre la misma razón doble, cualquiera que sea la posición de dicha recta. (La demostración de esta propiedad se deja al lector como ejercicio.) Podemos, pues, adoptar este valor como el de la razón doble de los cuatro planos. En forma equivalente, se puede definir la razón doble de cuatro planos coaxiales por medio de la de las cuatro rectas determinadas al cortarlos por un quinto plano (véase Fig. 79).

El concepto de razón doble de cuatro planos nos lleva naturalmente a la pregunta de si puede definirse una transformación proyectiva del espacio *tridimensional* en sí mismo. No es posible generalizar de modo inmediato la definición mediante una proyección central, de dos a tres dimensiones; pero puede demostrarse que toda transfor-

mación continua de un plano en sí mismo que establezca una correspondencia biunívoca entre puntos y puntos, por una parte, y rectas y rectas, por otra, es una transformación proyectiva. Este teorema sugiere la siguiente definición para el caso de tres dimensiones. Una

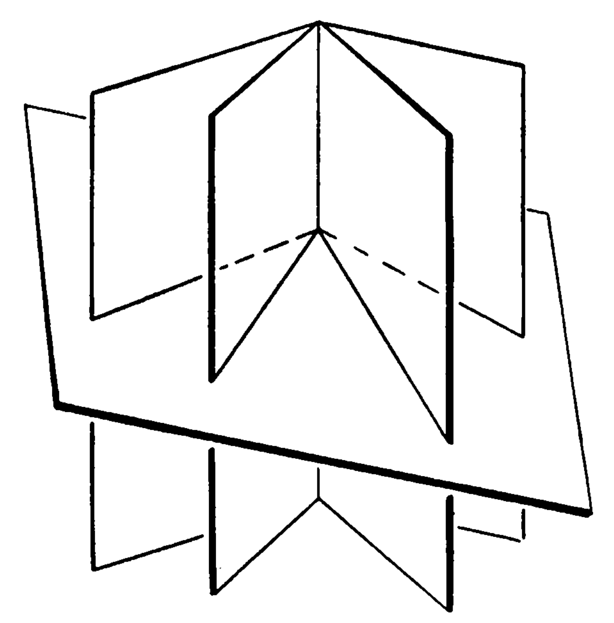


FIG. 79.—Razón doble de planos coaxiales.

transformación proyectiva del espacio es toda transformación biunívoca y continua que conserva las rectas. Puede demostrarse que en estas transformaciones la razón doble permanece invariante.

Las proposiciones precedentes se pueden completar mediante algunas observaciones. Supongamos dados tres puntos distintos A, B, C de una recta, con abscisas x_1, x_2, x_3; se desea encontrar un cuarto punto D, tal que $(ABCD) = \lambda$, siendo λ prefijado. (El caso especial $\lambda = -1$, para el cual el problema equivale a la construcción del cuarto armónico, será explicado con más detalle en la sección próxima.) En general, el problema tiene solución única, pues si es x la abscisa del punto pedido D, la ecuación

$$\frac{x_3 - x_1}{x_3 - x_2} \cdot \frac{x - x_2}{x - x_1} = \lambda \qquad [2]$$

tiene exactamente una solución x. Si se dan x_1, x_2, x_3, y simplificamos la ecuación [2] haciendo $(x_3 - x_1)/(x_3 - x_2) = k$, encontraremos, al

resolver dicha ecuación, que $x = (kx_2 - \lambda x_1)/(k - \lambda)$; p. ej., si los tres puntos A, B, C son equidistantes, de abscisas $x_1 = 0$, $x_2 = d$, $x_3 = 2d$, respectivamente, entonces: $k = (2d - 0)/(2d - d) = 2$, y $x = 2d/(2 - \lambda)$.

Si proyectamos la misma recta l sobre dos rectas distintas, l' y l'', desde dos centros diferentes O' y O'', obtenemos una correspondencia

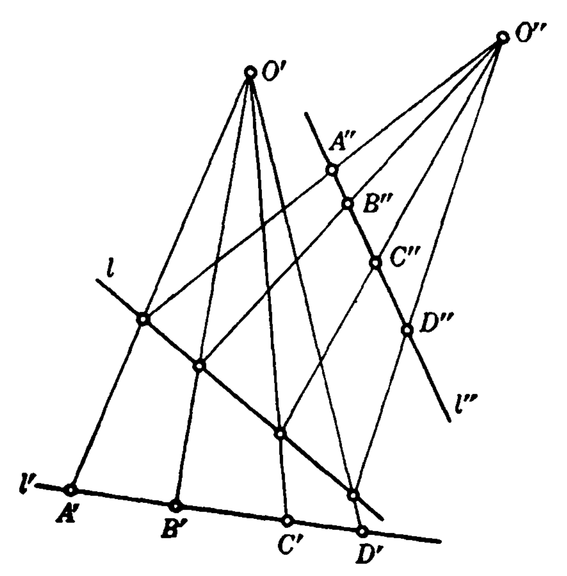


FIG. 80.—Correspondencia proyectiva entre los puntos de dos rectas.

$P \longleftrightarrow P'$ entre los puntos de l y l', y otra $P \longleftrightarrow P''$ entre los de l y l''. Esto establece una correspondencia $P' \longleftrightarrow P''$ entre los puntos de l' y los de l'', con la propiedad de que toda cuaterna de puntos A', B', C', D', de l' tiene la misma razón doble que la cuaterna correspondiente A'', B'', C'', D'', de l''. Toda correspondencia biunívoca entre los puntos de dos rectas que goce de esta propiedad se llamará *correspondencia proyectiva*, cualquiera que sea la forma en que haya sido definida.

Ejercicios:

1. Demuéstrese que, dadas dos rectas en correspondencia proyectiva, puede trasladarse una de ellas mediante un desplazamiento paralelo hasta una posición tal que la correspondencia dada se obtenga por simple proyección. (*Sugerencia:* háganse coincidir un par de puntos correspondientes de las dos rectas.)

2. Basándose en el resultado precedente, pruébese que si los puntos de las rectas l y l' se coordinan mediante una sucesión finita de proyecciones sobre varias

rectas intermedias, usando centros arbitrarios de proyección, el mismo resultado puede obtenerse mediante sólo *dos* proyecciones.

2. Aplicación al cuadrilátero completo.—Como aplicación interesante de la invariancia de la razón doble, estableceremos un teorema sencillo, pero importante, de geometría proyectiva. Se refiere al *cuadrilátero completo,* figura que consta de cuatro rectas cualesquiera (no concurrentes tres a tres), y de los seis puntos en que se cortan. En la figura 81, las cuatro rectas son AE, BE, BI y AF. Las rectas AB, EG e IF son las *diagonales* del cuadrilátero. Elijamos una diagonal, p. ej., AB, y señalemos sobre ella los puntos C y D donde encuentra a las otras dos diagonales. El teorema dice entonces que $(ABCD) = -1$, o sea: *los puntos de intersección de una diagonal con las otras dos separan armónicamente a los vértices del cuadrilátero situados sobre dicha diagonal.*

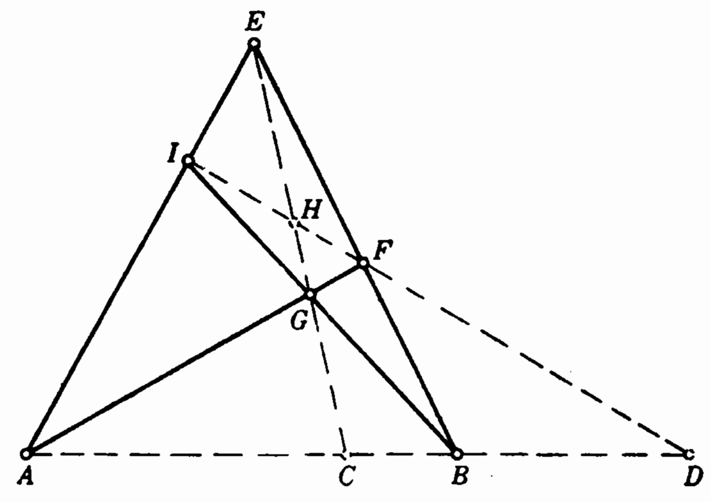


FIG. 81.—Cuadrilátero completo.

Para probarlo basta observar que

$$x = (ABCD) = (IFHD) \quad \text{por proyección desde } E.$$
$$(IFHD) = (BACD) \quad \text{por proyección desde } G.$$

Pero sabemos que $(BACD) = 1/(ABCD)$, o sea, $x = 1/x$, $x^2 = 1$, $x = \pm 1$, y como C, D separan a A y B, la razón doble x debe ser negativa; es decir, igual a -1, quedando demostrado el teorema.

Esta notable propiedad del cuadrilátero completo nos permite

construir, sin más ayuda que la regla, el conjugado armónico, respecto a un par A, B, de un tercer punto colineal C. Necesitamos solamente elegir un punto E fuera de la recta; trazar EA, EB, EC; elegir un punto G sobre EC; trazar AG y BG, que cortan a EB y EA en F e I, respectivamente, y dibujar IF, la cual corta a la recta ABC en el punto D, cuarto armónico pedido.



FIG. 82.—Prolongación de una recta al otro lado de un obstáculo.

Problema.—Dado un segmento AB en el plano y una región R, como se muestra en la figura 82, se desea prolongar la recta a la derecha de R. ¿Cómo puede lograrse esto con la regla sola, sin atravesar R en la construcción? (*Sugerencia:* Elíjanse dos puntos C, C' del segmento AB, y hállense sus conjugados armónicos D, D', respectivamente, por medio de cuatro cuadriláteros que tengan A y B como vértices.)

IV. PARALELISMO E INFINITO

1. Puntos del infinito como «puntos ideales».—Algunos de los argumentos utilizados en la sección precedente fallan si ciertas rectas de las utilizadas en las construcciones, supuestas prolongadas hasta su intersección, resultan paralelas; p. ej., en la construcción del cuarto armónico, D deja de existir si la recta IF resulta paralela a AB. El razonamiento geométrico parece complicarse a cada paso por el hecho de que dos rectas paralelas no se cortan, ya que en toda discusión que incluya la intersección de rectas el caso excepcional de paralelismo debe considerarse y formularse por separado. Asimismo, la proyección desde un centro O debe distinguirse de la proyección paralela, la cual exige tratamiento aparte. Si realmente entrásemos en una detallada discusión de cada caso excepcional, la geometría proyectiva resultaría muy complicada. Nos vemos inducidos, por tanto, a ensayar un nuevo recurso, *el de hallar generalizaciones de nuestros conceptos básicos, que eliminen las excepciones.*

Aquí la intuición geométrica nos indica el camino; si una recta que corta a otra gira lentamente hasta ser paralela, el punto de intersección de ambas se alejará hasta el infinito. Podemos simplemente decir que las dos rectas se cortan en un «punto del infinito». Lo esencial es, entonces, dar significado preciso a esta vaga afirmación, con el fin de poder operar con los puntos del infinito o, como a veces se dice, puntos ideales, del mismo modo que con los puntos ordinarios del plano o del espacio. En otras palabras, deseamos que todas las

reglas concernientes a las relaciones entre puntos, rectas y planos, etc., persistan aun cuando estos elementos geométricos sean ideales. Para alcanzar este propósito podemos proceder, bien intuitiva o formalmente, tal como se hizo al extender el sistema numérico, donde un método consistía en la idea intuitiva de medida, y el otro partía de las reglas formales de las operaciones aritméticas.

Antes de nada, observemos que en geometría sintética, incluso los conceptos básicos de punto y recta «ordinarios» no están definidos matemáticamente. Las llamadas definiciones de estos entes que encontramos con frecuencia en los textos de geometría elemental son sólo descripciones más o menos sugerentes. En el caso de los elementos geométricos ordinarios, nuestra intuición nos tranquiliza en lo que a su «existencia» se refiere. Pero lo que realmente necesitamos en geometría, considerada como un sistema matemático, es la validez de ciertas reglas mediante las cuales podamos operar con estos conceptos, tales como unir puntos, hallar la intersección de rectas, etc. Lógicamente considerado, un «punto» no es una «cosa en sí», sino que está caracterizado por la totalidad de las proposiciones mediante las cuales se halla relacionado con otros objetos. La existencia matemática de «puntos del infinito» quedará asegurada en cuanto se establezcan, de forma clara y no contradictoria, las *propiedades* matemáticas de estos nuevos entes; es decir, sus relaciones con los puntos «ordinarios» y entre sí. Los axiomas ordinarios de la geometría (p. ej., el de Euclides) son abstracciones del mundo físico, de señales de lápiz o de tiza, cuerdas tensas, rayos de luz, varillas rígidas, etc. Las propiedades que estos axiomas atribuyen a los puntos y rectas matemáticos son descripciones muy simplificadas e idealizadas de la conducta de sus representantes físicos. Entre dos señales puntiformes marcadas con lápiz, pueden trazarse no una, sino varias rectas. Si los puntos se hacen cada vez más pequeños, entonces todas estas rectas tendrán aproximadamente la misma apariencia. Esto es lo que se piensa cuando se establece como axioma geométrico que «por dos puntos puede trazarse *una* recta y *sólo una*»; no nos referimos a los puntos y rectas físicos, sino a los puntos y rectas conceptuales y abstractos de la geometría. Los puntos y rectas geométricos tienen propiedades esencialmente más sencillas que las de los objetos físicos, y esta simplificación procura la condición esencial para el desarrollo de la geometría como ciencia deductiva.

Conforme hemos señalado, la geometría ordinaria de los puntos y rectas se complica grandemente por el hecho de que un par de rectas paralelas no tienen punto de intersección. Nos vemos, pues, obliga-

dos a introducir una posterior simplificación en la estructura de la geometría, ampliando el concepto de punto geométrico, para hacer desaparecer esta excepción. Y, análogamente a la forma en que hemos ampliado el concepto de número para hacer posibles sin restricción la resta y la división, aquí también nos guiará el deseo de conservar, en el dominio ampliado, las leyes que gobiernan el dominio original.

Convenimos, pues, en agregar a los puntos ordinarios de cada recta un solo punto «ideal». Este punto se considerará como perteneciente a todas las rectas paralelas a la dada y únicamente a ellas. Como consecuencia de este convenio, *todo* par de rectas del plano se cortará en un solo punto; si las rectas no son paralelas, su intersección será un punto ordinario, mientras que si lo son, se cortarán en el punto ideal común a las dos rectas. Por razones intuitivas, el punto ideal de una recta se llama *punto del infinito* de la misma.

El concepto intuitivo de punto en el infinito de una recta parece sugerir que debían agregarse *dos* puntos ideales a cada recta. La razón para que sea uno solo, tal como hemos hecho, es que deseamos conservar la ley de que por dos puntos puede trazarse una y sólo una recta. Si una recta tuviera dos puntos del infinito comunes con toda recta paralela, entonces por estos dos «puntos» pasarían infinitas rectas paralelas.

Convendremos también en agregar a las rectas ordinarias del plano una sola recta «ideal» (llamada también recta del infinito del plano), *la cual contiene todos los puntos ideales del plano y ningún otro punto.* Se nos impone precisamente este convenio si deseamos conservar la ley inicial de que entre dos puntos cualesquiera puede trazarse una recta y la nueva ley, según la cual dos rectas cualesquiera se cortan en un punto. Para ver esto, elijamos dos puntos ideales arbitrarios. La recta única determinada por estos dos puntos no puede ser una recta ordinaria, ya que por el convenio adoptado, toda recta ordinaria contiene un solo punto ideal. Además, esta recta no puede contener puntos ordinarios, puesto que un punto ordinario y uno ideal determinan una recta ordinaria. Finalmente, esta recta debe contener *todos* los puntos ideales, si deseamos que tenga un punto común con toda recta ordinaria. En consecuencia, esta recta debe tener precisamente las propiedades que hemos asignado a la recta ideal del plano.

De acuerdo con nuestros convenios, un punto del infinito está determinado o representado por una familia de rectas paralelas, así como un número irracional viene determinado por una sucesión de intervalos racionales. La afirmación de que la intersección de dos rectas paralelas es un punto del infinito carece de significado misterioso; es sólo una manera cómoda de afirmar que las rectas son paralelas.

Esta forma de expresar el paralelismo en un lenguaje originalmente reservado para objetos intuitivamente diferentes, tiene el único propósito de evitar la enumeración de los casos excepcionales superfluos; éstos quedan ahora automáticamente incluídos en las mismas expresiones verbales u otros símbolos utilizados para los casos «ordinarios».

En resumen, nuestros convenios acerca de los puntos del infinito han sido elegidos de manera que las leyes que rigen la relación de incidencia entre puntos y rectas ordinarias sigan cumpliéndose en el dominio ampliado de puntos, mientras que la operación de hallar el punto de intersección de dos rectas, antes sólo posible si éstas no eran paralelas, puede ahora efectuarse sin ninguna restricción. Las consideraciones que llevan a esta simplificación formal en las propiedades de la relación de incidencia pueden parecer algo abstractas, pero quedan ampliamente justificadas por el resultado, como el lector comprobará en las páginas que siguen.

2. Elementos ideales y proyección.—La introducción de los puntos y la recta del infinito en el plano nos permite tratar la proyección

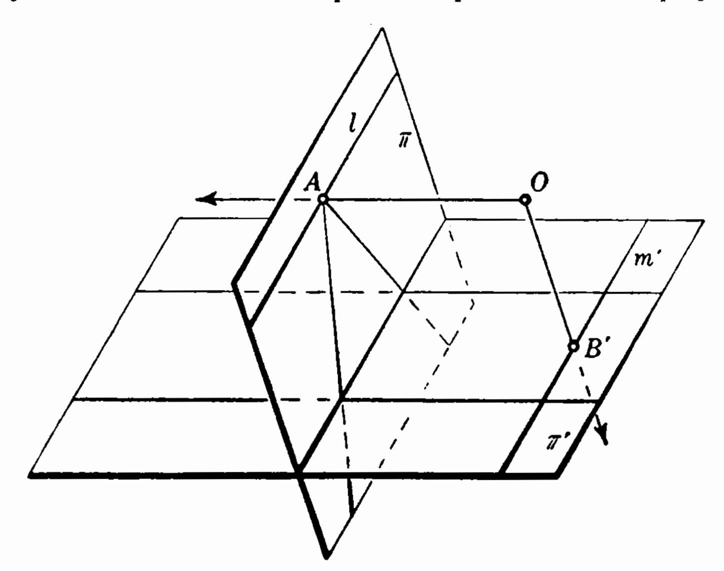


FIG. 83.—Proyección en elementos del infinito.

de un plano sobre otro en forma mucho más satisfactoria. Consideremos la proyección de un plano π sobre otro plano π', desde un centro O (Fig. 83). Esta proyección establece una correspondencia entre los puntos y las rectas de π y de π'. A cada punto A de π corresponde

un solo punto A' de π', con las siguientes excepciones: si el rayo proyectante que pasa por O es *paralelo* al plano π', entonces corta al plano π en un punto A, al que no corresponde ningún punto ordinario de π'. Estos puntos excepcionales de π están en una recta l a la cual no corresponde ninguna recta ordinaria de π'. Pero estas excepciones se eliminan, si convenimos en que a A corresponde el punto del infinito de π' en la dirección de la recta OA, y que a l corresponde la recta del infinito de π'. En igual forma, asignamos un punto del infinito de π a todo punto B', de la recta m' de π', por la que pasan todos los rayos trazados por O paralelos al plano π. A m' le corresponde la recta del infinito de π. De este modo, mediante la introducción de los puntos y la recta del infinito del plano, *una proyección de un plano sobre otro establece una correspondencia entre los puntos y rectas de ambos planos, que es biunívoca sin excepción.* (Esto hace desaparecer las excepciones mencionadas en la nota [2] de la pág. 181.) Además, es fácil ver que, como consecuencia de nuestro convenio, *un punto estará sobre una recta si, y sólo si, la proyección del punto está sobre la proyección de la recta.* Por tanto, todas las proposiciones acerca de puntos colineales, rectas concurrentes, etc., que incluyan sólo puntos, rectas y la relación de incidencia, deben considerarse como invariantes respecto a la proyección en sentido amplio. Esto nos permite tratar de los puntos del infinito de un plano π operando con los puntos ordinarios correspondientes de un plano π', relacionado con π mediante una proyección.

*La interpretación de los puntos del infinito de un plano π, por medio de una proyección desde un punto exterior O sobre puntos ordinarios de otro plano π', puede utilizarse para dar un «modelo» euclídeo concreto del plano ampliado. Para este fin, olvidémonos del plano π' y fijemos nuestra atención sobre π y las rectas que pasan por O. A cada punto ordinario de π corresponde una recta que pasa por O y no es paralela a π; a cada punto del infinito de π corresponde una recta que pasa por O y es paralela a π. Luego a la totalidad de los puntos ordinarios e ideales de π, corresponde la totalidad de las rectas que pasan por O, y esta correspondencia es biunívoca sin excepción. A los *puntos de una recta* de π corresponden las *rectas* de un *plano* que pasa por O. Un punto y una recta de π son incidentes, si, y sólo si, la recta y plano correspondientes que pasan por O lo son. Por tanto, la geometría de la incidencia de puntos y rectas del plano ampliado es por entero equivalente a la geometría de la incidencia de rectas y planos ordinarios que pasan por un punto fijo del espacio.

*En tres dimensiones la situación es similar, aunque ya no pode-

mos concretarla intuitivamente por proyección. De nuevo introducimos un punto del infinito asociado con cada familia de rectas paralelas. En cada plano tendremos una recta del infinito. A continuación deberemos introducir un nuevo elemento, el *plano del infinito*, formado por todos los puntos del infinito del espacio y que contiene todas las rectas del infinito. Cada plano ordinario corta al plano del infinito en su recta del infinito.

3. Razón doble con elementos en el infinito.—Debe hacerse una aclaración acerca de la razón doble cuando incluye elementos del infinito. Designemos el punto del infinito de una recta l por el símbolo ∞. Si A, B, C son tres puntos ordinarios de l, entonces podemos asignar un valor al símbolo $(ABC\infty)$ de la siguiente manera: elegido un punto P de l, $(ABC\infty)$ será el límite a que tiende $(ABCP)$ cuando P se aleja infinitamente sobre l; pero

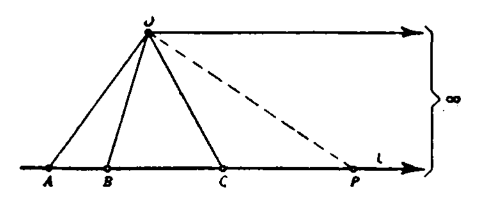


FIG. 84.—Razón doble con un punto en el infinito.

$$(ABCP) = \frac{CA}{CB} : \frac{PA}{PB},$$

y cuando P se aleja al infinito, PA/PB tiende a 1; en consecuencia, definimos $(ABC\infty) = CA/CB$.

En particular, si $(ABC\infty) = -1$, C es el punto medio del segmento AB. *El punto medio de un segmento y el punto del infinito de la recta soporte dividen armónicamente al segmento.*

Ejercicios: ¿Cuál es la razón doble de las cuatro rectas l_1, l_2, l_3, l_4, si son paralelas? ¿Cuál es la razón doble, si l_4 es la recta del infinito?

V. APLICACIONES

1. Notas preliminares.—Con la introducción de los elementos del infinito ya no es necesario enunciar explícitamente los casos excepcionales que surgen en las construcciones y teoremas, cuando dos o más rectas son paralelas. Sólo necesitamos recordar que si un punto está en el infinito, todas las rectas que pasan por él son paralelas. La distinción entre proyección central y paralela resulta innecesaria, pues la última significa simplemente proyección desde un punto del infinito. En la figura 72 el punto O o la recta PQR pueden ser del infinito

(la Fig. 85 muestra el primer caso). Se deja como ejercicio al lector el formular en lenguaje «finito» el correspondiente enunciado del teorema de Desargues.

No sólo el *enunciado,* sino también la *demostración* de un teorema proyectivo resulta con frecuencia más sencilla mediante el uso de elementos del infinito. El principio general es el siguiente: por «clase proyectiva» de una figura geométrica F entendemos el conjunto de todas las figuras que pueden resultar de F por medio de transformaciones proyectivas. Las propiedades proyectivas de F serán idénticas a las de cualquier miembro de su clase proyectiva, ya que las

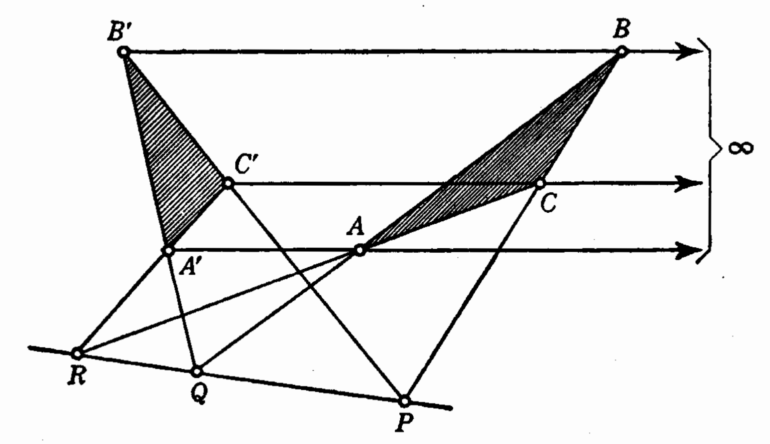


FIG. 85.—Configuración de Desargues con el centro en el infinito.

propiedades proyectivas son, por definición, invariantes en la proyección. Así, todo teorema proyectivo (es decir, referente sólo a propiedades proyectivas) que se verifique para F será cierto también para todo miembro de la clase proyectiva de F, y recíprocamente. Por tanto, para demostrar uno de tales teoremas para F, bastará hacerlo para otro miembro cualquiera de su clase proyectiva. Es posible a veces sacar ventaja de esto hallando un miembro especial de la clase proyectiva de F para el cual el teorema sea más fácil de demostrar que en el caso de la propia figura F; p. ej., dos puntos cualesquiera A, B de un plano π pueden transformarse en puntos del infinito, proyectándolos desde un centro O sobre un plano π' paralelo al plano de OAB; las rectas que pasan por A, y aquellas que pasan por B, se transformarán en dos familias de rectas paralelas. En los teoremas proyectivos que vamos a demostrar en esta sección haremos uso de esta transformación preliminar.

La siguiente propiedad elemental referente a rectas paralelas nos será muy útil en lo que sigue: sean dos rectas que pasan por un

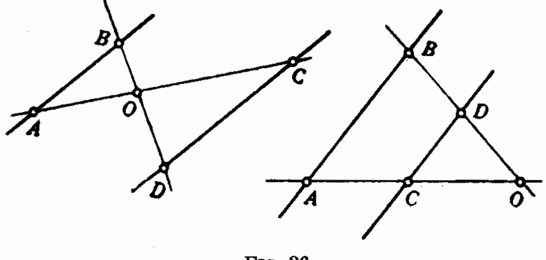


Fig. 86.

punto O, y cortan a un par de rectas l_1 y l_2 en los puntos A, B, C, D, tal como se ve en la figura 86. Si l_1 y l_2 son paralelas,

$$\frac{OA}{OC} = \frac{OB}{OD}.$$

y, recíprocamente, si $\frac{OA}{OC} = \frac{OB}{OD}$, l_1 y l_2 son paralelas. La demostración es consecuencia inmediata de propiedades elementales de la semejanza de triángulos y la dejamos como ejercicio al lector.

2. Demostración del teorema de Desargues en el plano.—Vamos a dar ahora la demostración de que los triángulos ABC y $A'B'C'$ de un plano, situados como en la figura 72, en los que las rectas determi-

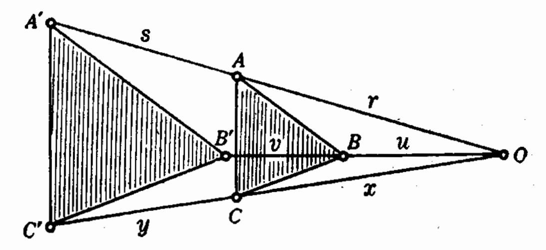


Fig. 87.—Demostración del teorema de Desargues.

nadas por vértices correspondientes concurren en un mismo punto, tienen las intersecciones P, Q y R de los lados correspondientes en línea recta. Para ello, proyectemos primero la figura de modo que Q y R pasen a ser puntos del infinito. Después de la proyección, AB será paralela a $A'B'$ y AC lo será a $A'C'$, con lo cual la figura aparecerá según se ve en la figura 87. Como ya antes hemos indicado, para

demostrar el teorema de Desargues en general, basta hacerlo sobre este tipo especial de figura. Para nuestro objeto necesitamos sólo probar que la intersección de BC y $B'C'$ es también un punto del infinito; es decir, que BC es paralela a $B'C'$; entonces, P, Q y R serán colineales (ya que estarán en la recta del infinito). Ahora bien:

$$AB \parallel A'B' \text{ implica } \frac{u}{v} = \frac{r}{s},$$

y

$$AC \parallel A'C' \text{ implica } \frac{x}{y} = \frac{r}{s}$$

Por consiguiente, $u/v = x/y$, lo que supone $BC \parallel B'C'$, como queríamos probar.

Observemos que en esta demostración del teorema de Desargues hacemos uso de la noción de longitud de un segmento; es decir, que hemos demostrado un teorema proyectivo con medios métricos. Además, si las transformaciones proyectivas se definen «intrínsecamente» como transformaciones planas que conservan la razón doble (véase pág. 189), entonces la demostración se efectúa por completo en el plano.

Ejercicio: Demuéstrese, en forma análoga, el recíproco del teorema de Desargues: si los triángulos ABC y $A'B'C'$ tienen la propiedad de que P, Q y R son colineales, las rectas AA', BB' y CC' son concurrentes.

3. Teorema de Pascal [1].—Este teorema dice: *si los vértices de un hexágono están alternativamente sobre un par de rectas concurrentes, las tres intersecciones* P, Q *y* R *de los lados opuestos son colineales* (Fig. 88). (El hexágono puede cortarse a sí mismo.) Los lados «opuestos» se identifican en el diagrama esquemático de la figura 89.

Mediante una proyección preliminar, supondremos que P y Q son puntos del infinito. Necesitamos probar, pues, que también R está en el infinito. La disposición está aclarada en la figura 90, en la que $23 \parallel 56$ y $12 \parallel 45$. Para probar que $16 \parallel 34$, tenemos

$$\frac{a}{a+x} = \frac{b+y}{b+y+s}, \qquad \frac{b}{b+y} = \frac{a+x}{a+x+r}$$

Por consiguiente,

$$\frac{a}{b} = \frac{a+x+r}{b+y+s},$$

es decir, $16 \parallel 34$, como queríamos demostrar.

[1] Más adelante (pág. 221) discutiremos un teorema más general del mismo tipo. El presente caso especial se conoce también por el nombre de su descubridor, Pappus de Alejandría (siglo III a. de J. C.).

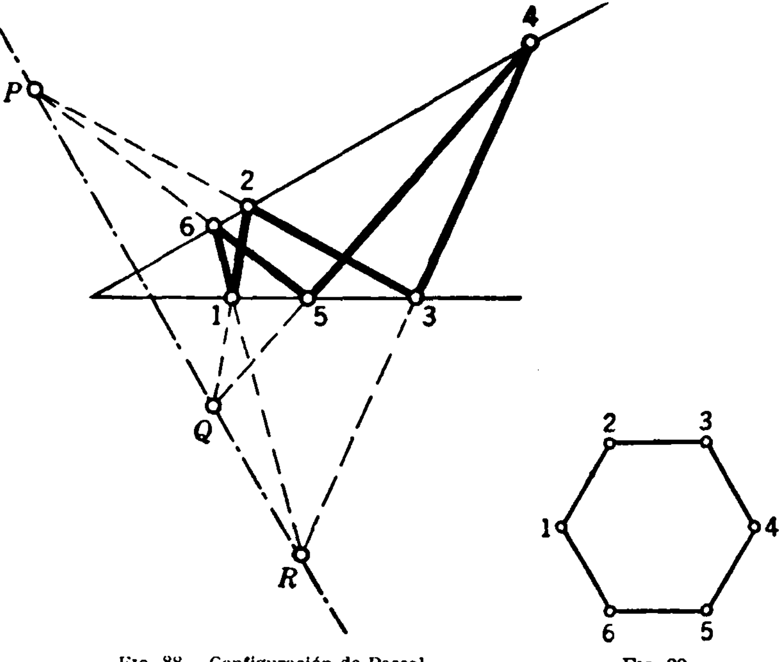


Fig. 88.—Configuración de Pascal.

Fig. 89.

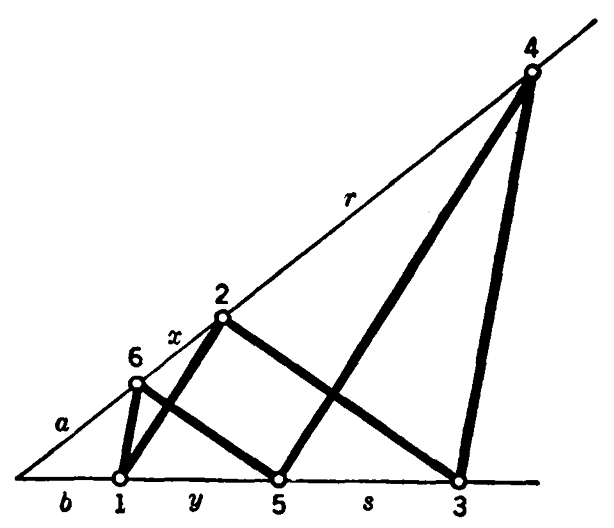


Fig. 90.—Demostración del teorema de Pascal.

4. Teorema de Brianchon.—Este teorema dice: *si los lados de un hexágono pasan alternativamente por dos puntos fijos* P *y* Q, *las tres diagonales que unen pares de vértices opuestos del hexágono son concu-*

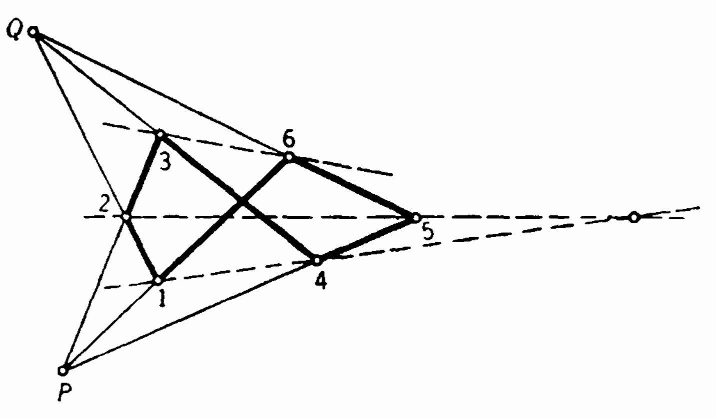


Fig. 91.—Configuración de Brianchon.

rrentes (véase Fig. 91). Mediante una proyección podemos llevar al infinito el punto P y el punto en que se cortan dos diagonales, p. ej., 14 y 36 (véase Fig. 92). Como 14 $\parallel$ 36, se tiene $a/b = u/v$. Pero $x/y = a/b$ y $u/v = r/s$; por consiguiente, $x/y = r/s$ y 36 $\parallel$ 25; es decir, las tres

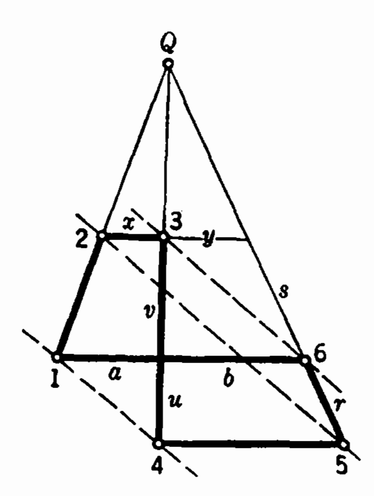


Fig. 92.—Demostración del teorema de Brianchon.

diagonales son paralelas y, por tanto, concurrentes. Esto basta para probar el teorema en el caso general.

5. Nota sobre la ley de dualidad.—El lector habrá observado la notable analogía entre los teoremas de Pascal (1623-1662) y de Brianchon (1785-1864). Esta semejanza resulta particularmente sorprendente si escribimos los teoremas uno al lado del otro.

TEOREMA DE PASCAL	TEOREMA DE BRIANCHON
Si los *vértices* de un hexágono *están alternativamente sobre dos rectas*, los *puntos en que se cortan los lados opuestos* son *colineales.*	Si los *lados* de un hexágono *pasan alternativamente por* dos *puntos*, las *rectas que unen vértices opuestos* son *concurrentes.*

No sólo los teoremas de Pascal y Brianchon, sino todos los teoremas de geometría proyectiva se presentan a pares, análogos uno a otro y, por decirlo así, idénticos en estructura. Esta relación se llama *dualidad.* En geometría plana el punto y la recta se denominan *elementos duales.* Trazar una recta que pase por un punto, o señalar un punto sobre una recta, son *operaciones duales.* Dos figuras son duales si una puede deducirse de la otra reemplazando cada elemento y operación por el elemento y operación duales; p. ej., los teoremas de Pascal y Brianchon son duales, y el dual del teorema de Desargues es precisamente su recíproco. Este fenómeno de la dualidad da a la geometría proyectiva un carácter muy distinto al que tiene la geometría métrica elemental, en la cual no existe tal dualidad (p. ej., carece de significado hablar del dual de un ángulo de 37° o de un segmento de longitud 2). En muchos textos de geometría proyectiva *el principio de dualidad,* que dice que *el dual de todo teorema verdadero de geometría proyectiva es asimismo un teorema verdadero de geometría proyectiva,* se pone de manifiesto colocando los teoremas duales junto con sus demostraciones duales en columnas paralelas de cada página, según hemos hecho anteriormente. La razón básica de esta dualidad será considerada en la sección siguiente (véase también pág. 229).

VI. REPRESENTACIÓN ANALÍTICA

1. Observaciones preliminares.—En el desarrollo inicial de la geometría proyectiva existió una acusada tendencia a construirlo todo sobre una base sintética y *puramente geométrica,* evitando el uso de números y métodos algebraicos. Este programa tropezó con grandes dificultades, ya que siempre quedaban puntos donde parecía inevitable

alguna formulación algebraica. Hacia los últimos años del siglo XIX se consiguió, no obstante, éxito completo en la construcción de una geometría proyectiva puramente sintética, pero a costa de gran complicación. En este sentido, los métodos de la geometría analítica han tenido mucho más éxito. La tendencia general en la matemática moderna es basarlo todo en el concepto de número; y, en geometría, esta tendencia, iniciada con Fermat y Descartes, ha obtenido triunfos decisivos. La geometría analítica se ha desarrollado desde el estadio de un mero instrumento de razonamiento geométrico hasta constituir una disciplina en la que la interpretación geométrica intuitiva de las operaciones y resultados no es ya el fin último y exclusivo, sino que tiene más bien la función de ser un principio director que ayuda a sugerir y comprender los resultados analíticos. Este cambio en el significado de la geometría es el producto de un desarrollo histórico gradual que ha ampliado enormemente el objeto de la geometría clásica, y que, al propio tiempo, ha producido una unión casi orgánica de la geometría y el análisis.

En geometría analítica, las «coordenadas» de un ente geométrico son un conjunto de números que caracterizan a dicho objeto unívocamente; p. ej., un punto se define dando sus coordenadas rectangulares x, y o sus coordenadas polares ρ, θ, mientras un triángulo puede definirse por las coordenadas de sus tres vértices, lo que requiere en total seis números. Sabemos que una recta en el plano x, y es el lugar geométrico de todos los puntos $P(x, y)$ (véase pág. 83 para esta notación) cuyas coordenadas satisfacen a una ecuación lineal:

$$ax + by + c = 0. \qquad [1]$$

Podremos, por consiguiente, designar los tres números a, b, c como las «coordenadas» de esta recta; p. ej., $a = 0$, $b = 1$, $c = 0$ definen la recta $y = 0$, que es el eje x; $a = 1$, $b = 1$, $c = 0$ definen la recta $x = y$, que es la bisectriz del primer cuadrante. Del mismo modo, las ecuaciones cuadráticas definen «secciones cónicas»:

$x^2 + y^2 = r^2$ circunferencia de centro en el origen y radio r;

$(x - a)^2 + (y - b)^2 = r^2$ circunferencia de centro en (a, b) y radio r;

$\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$ elipse,

etcétera.

El método que se ofrece como más natural en geometría analítica es partir de conceptos puramente «geométricos»—puntos, rectas, etc.—

y traducirlos al lenguaje de los números. El punto de vista moderno es el recíproco: partimos del *conjunto de todos los pares de números* (x, y), y *llamamos* punto a cada uno de dichos pares, ya que podemos, si es nuestro deseo, *interpretar* cada par de números por la noción familiar de punto geométrico. Análogamente, una ecuación lineal entre x e y se dice que define una recta. El haber transferido la atención principal del aspecto intuitivo de la geometría al analítico abre el camino para un tratamiento sencillo, aunque riguroso, de los puntos del infinito de la geometría proyectiva, lo que es indispensable para una comprensión más profunda y completa de todo el tema. Para aquellos lectores que posean suficientes conocimientos preliminares, daremos una breve descripción de este método.

***2. Coordenadas homogéneas. Fundamento algebraico de la dualidad.**—En geometría analítica ordinaria, las coordenadas rectangulares de un punto del plano son sus distancias (con su correspondiente signo) a un par de ejes perpendiculares. El sistema falla cuando se consideran los puntos del infinito del plano ampliado de la geometría proyectiva. Por tanto, si deseamos aplicar los métodos analíticos a la geometría proyectiva, es necesario idear un sistema de coordenadas que comprenda tanto a los puntos ideales como a los ordinarios. La introducción de tal sistema de coordenadas se describe mejor suponiendo que el plano X, $Y(\pi)$ dado está «sumergido» en un espacio tridimensional, en el que se ha definido un sistema de coordenadas rectangulares x, y, z (o distancias afectadas de signo de un punto a los tres planos coordenados determinados por los ejes x, y, z). Coloquemos π paralelamente al plano coordenado x, y, y a una distancia 1 por encima de él, de forma que todo punto P de π tendrá coordenadas tridimensionales $(X, Y, 1)$. Tomando el origen O del sistema de coordenadas como centro de proyección, observemos que *cada punto* P *determina una única recta que pasa por* O, *y recíprocamente* (véase página 196. Las rectas que pasan por O y son paralelas a π corresponden a los puntos del infinito de π).

Vamos a describir ahora un sistema de «coordenadas homogéneas» para los puntos de π. Para hallar las coordenadas homogéneas de un punto ordinario P de π, consideramos la recta que une O con P y elegimos en ella *un* punto Q distinto del O (véase la Fig. 93). Entonces, las coordenadas tridimensionales x, y, z de Q son las *coordenadas homogéneas* de P. En particular, las coordenadas $(X, Y, 1)$ de P mismo constituyen un conjunto de coordenadas homogéneas de P. Además, cualquier otro conjunto de números (tX, tY, t) con $t \neq 0$ será también un conjunto de coordenadas homogéneas de P, ya que las coordenadas

de todos los puntos de la recta OP, excepto el O, serán de esta forma. [Excluímos el punto (0, 0, 0), que por estar en todas las rectas que pasan por O, no sirve para distinguir una de otra.]

Este método de definir coordenadas en el plano requiere tres números en lugar de dos para determinar la posición de un punto, y tiene la desventaja adicional de que las coordenadas de un punto no están determinadas unívocamente, sino que incluyen un factor arbitrario t. Pero ofrece la gran ventaja de que los puntos del infinito de π están ahora incluídos en la representación coordenada. Un punto P del infinito de π se determina mediante una recta que pasa por O

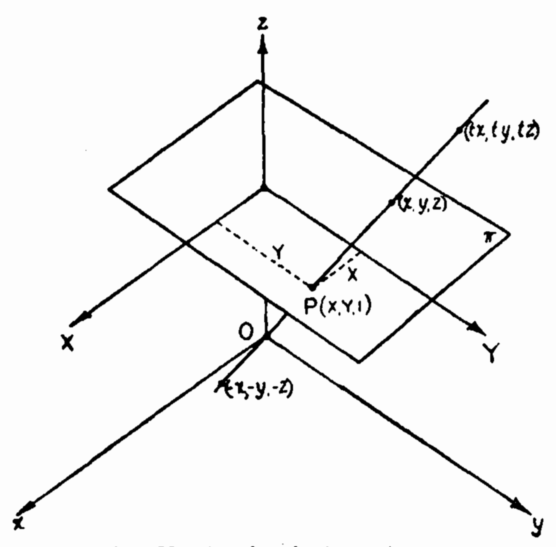


Fig. 93.—Coordenadas homogéneas.

y es paralela a π. Todo punto Q de esta recta tendrá coordenadas de la forma $(x, y, 0)$; por consiguiente, *las coordenadas homogéneas de un punto del infinito de* π *son de la forma* (x, y, 0).

La ecuación en coordenadas homogéneas de una recta de π puede hallarse inmediatamente, observando que las rectas que unen O con los puntos de esta recta están en un plano que pasa por O. Se demuestra en geometría analítica que la ecuación de tal plano es de la forma

$$ax + by + cz = 0.$$

Por tanto, esta es la ecuación en coordenadas homogéneas de una recta de π.

Ahora que el modelo geométrico de los puntos de π como rectas

que pasan por O ha cumplido su propósito, podemos dejarlo de lado y dar la siguiente definición, puramente analítica, del plano ampliado:

Un *punto* es una terna ordenada de números reales (x, y, z), no todos iguales a cero. Dos de tales ternas (x_1, y_1, z_1) y (x_2, y_2, z_2) definen el *mismo* punto, si para algún $t \neq 0$ es:

$$x_2 = tx_1 ,$$
$$y_2 = ty_1 ,$$
$$z_2 = tz_1 .$$

En otras palabras, las coordenadas de un punto cualquiera pueden multiplicarse por un factor no nulo sin que varíe el punto. (Por esta razón se llaman coordenadas *homogéneas*.) Un punto (x, y, z) es un punto *ordinario* si $z \neq 0$; si $z = 0$, es un *punto del infinito*.

Una *línea recta* de π consta de todos los puntos (x, y, z) que satisfacen a una ecuación lineal de la forma

$$ax + by + cz = 0, \qquad [1']$$

siendo a, b, c tres constantes no todas nulas. En particular, los puntos del infinito de π satisfacen todos a la ecuación lineal

$$z = 0. \qquad [2]$$

Ésta es, por definición, una recta llamada *recta del infinito* de π. Como una recta se define por una ecuación de la forma [1′], podemos llamar a la terna de números (a, b, c) *coordenadas homogéneas de la recta* [1′]. Sigue de ello que (ta, tb, tc), para cualquier $t \neq 0$, son también las coordenadas de la recta [1′], ya que la ecuación

$$(ta)x + (tb)y + (tc)z = 0 \qquad [3]$$

se satisface para las mismas ternas de coordenadas (x, y, z) que [1′].

En estas definiciones observamos una simetría perfecta entre punto y recta; cada uno de estos elementos está determinado por tres coordenadas homogéneas (u, v, w). La condición para que el punto (x, y, z) esté en la recta (a, b, c) es que

$$ax + by + cz = 0,$$

y ésta es asimismo la condición para que el punto cuyas coordenadas son (a, b, c) esté en la recta de coordenadas (x, y, z); p. ej., la identidad aritmética

$$2 \cdot 3 + 1 \cdot 4 - 5 \cdot 2 = 0$$

puede interpretarse tanto como que el punto (3, 4, 2) está en la recta (2, 1, −5) como que el punto (2, 1, −5) está en la recta (3, 4, 2). Esta simetría es la base de la dualidad entre punto y recta en geometría proyectiva, pues toda relación entre puntos y rectas se transforma en otra entre rectas y puntos cuando las coordenadas son adecuadamente reinterpretadas. En la nueva interpretación, las coordenadas anteriores de puntos y rectas deberán imaginarse ahora como representando rectas y puntos, respectivamente. Todas las operaciones algebraicas y resultados se conservan, pero su interpretación da la réplica dual del teorema original. Merece destacarse el hecho de que esta dualidad no se cumple en el plano ordinario de dos coordenadas X, Y, ya que la ecuación de una recta en coordenadas ordinarias

$$aX + bY + c = 0$$

no es simétrica respecto a X, Y y a, b, c. Sólo mediante la introducción de los puntos y de la recta del infinito queda perfectamente establecido el principio de dualidad.

Para pasar de las coordenadas homogéneas x, y, z de un punto ordinario P del plano π a las coordenadas rectangulares ordinarias, hacemos simplemente $X = x/z$, $Y = y/z$. Entonces, X, Y representan las distancias desde el punto P a dos ejes perpendiculares de π, paralelos a los ejes x e y, según se indica en la figura 93. Sabemos que una ecuación de la forma

$$aX + bY + c = 0$$

representa una recta de π. Al efectuar la sustitución $X = x/z$, $Y = y/z$ y multiplicar por z encontramos que la ecuación de la misma recta en coordenadas homogéneas es, como se ha dicho en la página anterior,

$$ax + by + cz = 0.$$

Por ejemplo, la ecuación de la recta $2x - 3y + z = 0$ es, en coordenadas rectangulares ordinarias, X, Y, $2X - 3Y + 1 = 0$. Naturalmente, la última ecuación falla para el punto del infinito de la recta, una de cuyas ternas de coordenadas homogéneas es (3, 2, 0).

Queda aún algo por decir: hemos conseguido dar una definición puramente analítica de punto y recta; pero ¿cómo obtener el concepto igualmente importante de transformación proyectiva? Puede demostrarse que una transformación proyectiva de un plano sobre otro, tal como se definió en la página 181, está dada analíticamente por un sistema de ecuaciones lineales

$$\begin{aligned} x' &= a_1x + b_1y + c_1z, \\ y' &= a_2x + b_2y + c_2z, \\ z' &= a_3x + b_3y + c_3z, \end{aligned} \qquad [4]$$

que relaciona las coordenadas homogéneas x', y', z' de los puntos del plano π' con las coordenadas homogéneas x, y, z de los puntos del plano π. Desde nuestro

punto de vista actual, podemos ahora *definir* una transformación proyectiva mediante un sistema de ecuaciones lineales de la forma [4]. Los teoremas de la geometría proyectiva se transforman, por consiguiente, en teoremas sobre el comportamiento de las ternas de números (x, y, z) bajo tales transformaciones; p. ej., la demostración de que la razón doble de cuatro puntos de una recta no varía al aplicar dichas transformaciones, es simplemente un ejercicio del álgebra de las transformaciones lineales. No entramos en los detalles de este procedimiento analítico, y volveremos, en cambio, a los aspectos más intuitivos de la geometría proyectiva.

VII. PROBLEMAS DE CONSTRUCCIÓN CON LA REGLA

En las construcciones siguientes se entiende que sólo se admite la regla como instrumento.

Los problemas 1 a 18 están contenidos en una memoria de J. Steiner, en la cual prueba que puede prescindirse del compás en las construcciones geométricas cuando se da una circunferencia fija y su centro (véase Cap. III, pág. 163). Se aconseja al lector que resuelva los problemas en el orden dado.

Una cuaterna de rectas a, b, c, d que pasan por un punto P se llama *armónica* si su razón doble $(abcd)$ es igual a -1; a y b se dicen *conjugados* de c y d, y viceversa.

1. Demuéstrese que si en una cuaterna armónica de rectas a, b, c, d, el rayo a es bisectriz del ángulo formado por c y d, b es entonces perpendicular a a.

2. Constrúyase el cuarto rayo armónico de una terna dada de rectas que pasan por un punto. (*Indicación:* Hágase uso del teorema del cuadrilátero completo.)

3. Constrúyase el cuarto punto armónico de una terna de puntos colineales dados.

4. Se dan un ángulo recto y un ángulo arbitrario que tienen el vértice y un lado comunes; duplíquese este ángulo.

5. Se da un ángulo y su bisectriz b. Constrúyase una perpendicular a b por el vértice del ángulo dado.

6. Demuéstrese que si las rectas $l_1, l_2, l_3, \ldots, l_n$, que pasan por un punto P, cortan a la recta a en los puntos $A_1, A_2, \ldots, A_n$ y a la recta b en los $B_1, B_2, \ldots, B_n$, todas las intersecciones de los pares de rectas A_iB_k y A_kB_i $(i \neq k; i,k = 1, 2, \ldots, n)$ están sobre una recta.

7. Demuéstrese que si una paralela al lado BC del triángulo ABC corta a AB en B' y a AC en C', la recta que une A con la intersección D de $B'C$ y $C'B$ corta a BC en su punto medio.

7*a*. Formúlese el recíproco de 7.

8. Sobre una recta l se dan tres puntos, P, Q, R, tales que Q es el punto medio del segmento PR. Constrúyase una paralela a l que pase por un punto dado S.

9. Dadas dos rectas paralelas, l_1 y l_2, hállese el punto medio de un segmento AB de l_1.

10. Trácese una paralela a dos rectas paralelas dadas l_1 y l_2 y que pase por un punto dado P. (*Indicación:* Redúzcase 9 a 7 y úsese 8.)

11. Steiner da la solución siguiente al problema de duplicar un segmento rectilíneo AB, cuando se da una paralela l a AB: por un punto C, que no está en l ni en AB, se traza CA, que corta a l en A_1, y CB, que corta a l en B_1. Dibújese entonces (véase 10) una paralela a l que pase por C, la cual cortará a BA_1 en D. Si DB_1 encuentra a AB en E, $AE = 2 \cdot AB$. Demuéstrese esto último.

12. Divídase el segmento AB en n partes iguales, si se tiene trazada una paralela l a AB. (*Indicación:* Constrúyase primero el múltiplo n-ésimo de un segmento arbitrario de l, por medio de 11.)

13. Dado un paralelogramo $ABCD$, trácese por un punto P una paralela a una recta dada l. (*Indicación:* Aplíquese 10 al centro del paralelogramo y úsese 8.)

14. Dado un paralelogramo, multiplíquese un segmento dado por n. (*Indicación:* Úsense 13 y 11.)

15. Dado un paralelogramo, divídase un segmento dado en n partes.

16. Se da una circunferencia y su centro; trácese una paralela a una recta dada desde un punto dado. (*Indicación:* Utilícese 13.)

17. Dados un círculo y su centro, multiplíquese y divídase un segmento dado por n. (*Indicación:* Utilícese 13.)

18. Se da una circunferencia fija y su centro; trácese una perpendicular a una recta dada por un punto dado. (*Indicación:* Úsese un rectángulo inscrito en la circunferencia y que tenga dos lados paralelos a la recta dada; redúzcase a ejercicios anteriores.)

19. Utilizando los resultados de los problemas 1 a 18, ¿cuáles son los problemas básicos de construcción que pueden resolverse sin otro instrumento que una regla con dos bordes paralelos?

20. Dos rectas dadas, l_1 y l_2, se cortan en un punto P *fuera* de los límites del papel. Constrúyase la recta que une un punto Q con P. (*Indicación:* Complétese con los elementos dados la figura del teorema de Desargues para el plano, de forma que P y Q sean intersecciones de lados correspondientes de dos triángulos del teorema de Desargues.)

21. Constrúyase la recta que une dos puntos dados cuya distancia es mayor que la longitud de la regla utilizada (véase 20).

22. Dos puntos P y Q, fuera de los límites del papel, están determinados por dos pares de rectas l_1, l_2 y m_1, m_2, respectivamente. Dibújese el segmento de la recta PQ que cae dentro del papel. (*Indicación:* Para obtener un punto de PQ, complétese la figura del teorema de Desargues, de manera que un triángulo tenga dos lados sobre l_1 y m_1 y los correspondientes del otro estén sobre l_2 y m_2.)

23. Resuélvase 20 por medio del teorema de Pascal (pág. 200). (*Indicación:* Complétese la figura correspondiente al teorema de Pascal considerando l_1 y l_2 como un par de lados opuestos del hexágono y Q como intersección de otro par de lados opuestos.)

*24. Dos rectas por completo fuera del papel están determinadas cada una por dos pares de rectas que se cortan en puntos exteriores al papel. Determínese el punto de intersección por medio de otro par de rectas que pasen por él.

VIII. CÓNICAS Y CUÁDRICAS

1. Geometría métrica elemental de las cónicas.—Hasta ahora nos hemos ocupado exclusivamente de puntos, rectas, planos y figuras formadas por combinaciones de estos entes. Si la geometría proyectiva consistiera únicamente en el estudio de esas figuras «lineales», tendría relativamente poco interés. Es un hecho de importancia fundamental que la geometría proyectiva *no* se limita al estudio de las figuras lineales, sino que incluye también todo el campo de las sec-

ciones cónicas y sus generalizaciones a más de tres dimensiones. El estudio métrico de Apolonio de las tres secciones cónicas—elipse, hipérbola y parábola—fué uno de los grandes triunfos de la matemática de la antigüedad. Es difícil exagerar la importancia de las secciones cónicas tanto en la matemática pura como en la aplicada (p. ej., las órbitas de los planetas y las de los electrones en el átomo de hidrógeno son secciones cónicas). No debe extrañarnos que la teoría clásica griega de las secciones cónicas sea todavía parte indispensable de la enseñanza matemática, pero de ninguna manera puede considerarse la matemática griega como algo definitivo. Dos mil años más tarde se descubrieron las importantes propiedades proyectivas de las cónicas, y a pesar de su sencillez y belleza, la inercia académica ha impedido hasta ahora que se introduzcan en los programas de la segunda enseñanza.

Vamos a empezar recordando las definiciones métricas de las secciones cónicas. Existen varias de estas definiciones cuya equivalencia se demuestra en geometría elemental. La usual parte del concepto de *focos.* Se define una *elipse* como el lugar geométrico de todos los puntos P del plano tales que la suma de sus distancias, r_1 y r_2, a dos puntos fijos, F_1 y F_2, llamados focos, es constante. (Si ambos focos coinciden, la figura es una circunferencia.) La *hipérbola* se define como el lugar geométrico de los puntos P del plano tales que el valor absoluto de la diferencia $r_1 - r_2$ es igual a una constante prefijada. Y, finalmente, se define la *parábola* como el lugar geométrico de todos los puntos P del plano para los cuales la distancia r a un punto fijo F es igual a la distancia a una recta dada l.

En geometría analítica todas estas curvas pueden expresarse mediante ecuaciones de segundo grado entre las coordenadas x e y. No es difícil demostrar, recíprocamente, que cualquier curva definida analíticamente por una ecuación de segundo grado:

$$ax^2 + by^2 + cxy + dx + ey + f = 0,$$

es una recta, o un par de rectas, o una de las tres cónicas. Esto se demuestra, generalmente, introduciendo un sistema nuevo de coordenadas más adecuado, como puede verse en cualquier curso de geometría analítica.

Estas definiciones de las secciones cónicas son esencialmente métricas, puesto que hacen uso del concepto de «distancia». Pero existe otra definición que establece el lugar que ocupan las secciones cónicas en la geometría proyectiva. *Las secciones cónicas son simplemente proyecciones de una circunferencia sobre un plano.* Si proyectamos una circunferencia C desde un punto O, todas las rectas proyectantes for-

marán un doble cono indefinido, cuya intersección con un plano π será la proyección de C. Esta intersección será una elipse o una hipérbola, según que el plano corte a una o a ambas hojas del cono. El caso intermedio de la parábola aparece cuando π es paralelo a una de las rectas que pasan por O (Fig. 94).

Este cono de proyección no precisa ser circular recto con su vértice O en la perpendicular levantada sobre el centro del círculo C; puede ser también oblicuo. En cualquier caso, y por ahora lo aceptaremos sin demostración, la intersección de un cono con un plano

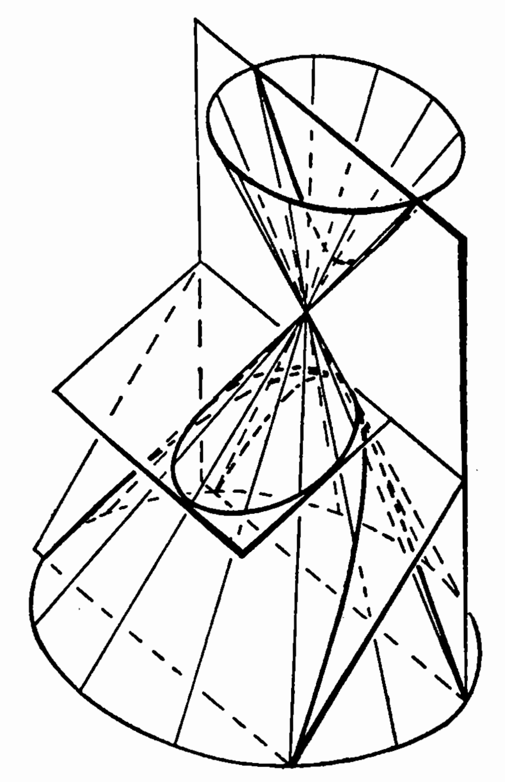


Fig. 94.—Secciones cónicas.

es una curva cuya ecuación es de segundo grado; recíprocamente, toda curva de segundo grado puede deducirse de una circunferencia por proyección. Por esta razón, las curvas de segundo grado se llaman secciones cónicas.

Hemos dicho que cuando el plano corta sólo a una hoja de un cono

circular recto, la curva de intersección E es una elipse. Podemos demostrar que E satisface la definición focal ordinaria de la elipse, tal como se ha dado antes, mediante una elegante argumentación debida al matemático belga G. P. Dandelin, quien la enunció en 1822. La

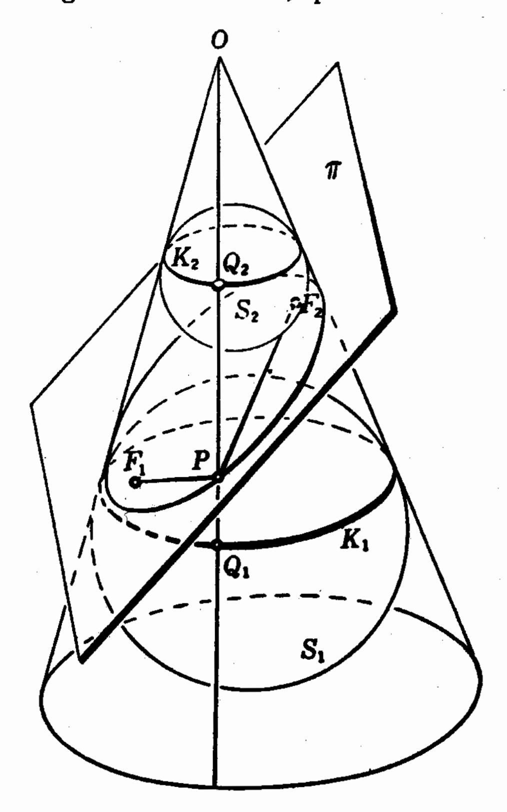


FIG. 95.—Esferas de Dandelin.

demostración se basa en la introducción de dos esferas S_1 y S_2 (Fig. 95), que son tangentes a π en los puntos F_1 y F_2, respectivamente, y tangentes al cono a lo largo de dos circunferencias situadas en planos paralelos K_1 y K_2, respectivamente. Unamos un punto cualquiera P de E con F_1 y F_2, y tracemos la recta que une P con el vértice O del cono. Esta recta es una generatriz de la superficie cónica y encuentra a las circunferencias K_1 y K_2 en los puntos Q_1 y Q_2, respectivamente. Como PF_1 y PQ_1 son dos tangentes desde P a S_1, se tiene:

$$PF_1 = PQ_1,$$

y, análogamente,

$$PF_2 = PQ_2 .$$

Sumando ambas igualdades miembro a miembro, resulta:

$$PF_1 + PF_2 = PQ_1 + PQ_2 = Q_1Q_2,$$

que es precisamente la distancia, contada sobre las generatrices del cono, entre los dos círculos paralelos K_1 y K_2, por lo cual resulta independiente de la particular elección del punto P sobre E. La ecuación resultante

$$PF_1 + PF_2 = \text{constante},$$

válida para todos los puntos P de E, es, precisamente, la definición focal de la elipse. En consecuencia, E es una elipse y F_1, F_2, sus focos.

Ejercicio: Cuando un plano corta las dos hojas de un cono, la curva de intersección es una hipérbola. Demuéstrese esta propiedad, utilizando una esfera en cada una de las hojas del cono.

2. Propiedades proyectivas de las cónicas.—Basándonos en los hechos acabados de exponer, adoptaremos la siguiente definición provisional: una cónica es la proyección de una circunferencia sobre un plano. Esta definición está más de acuerdo con el espíritu de la geometría proyectiva que la definición focal ordinaria, puesto que esta última se basa por entero en la noción métrica de distancia. Ni siquiera nuestra actual definición está libre de ese defecto, ya que la «circunferencia» es un concepto de geometría métrica. En seguida daremos una definición puramente proyectiva de las cónicas.

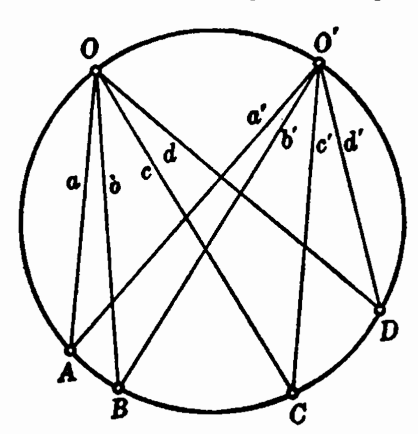


FIG. 96.—Razones dobles en una circunferencia.

Dado que nos hallamos de acuerdo con que una cónica es la proyección de una circunferencia (con la palabra «cónica» designamos cualquier curva de la clase proyectiva de la circunferencia; véase pág. 198), se deduce que cualquier propiedad de la circunferencia que permanezca invariante en la proyección será también una propiedad de las cónicas. Ahora bien: la

circunferencia tiene una propiedad (métrica) muy conocida, la de que un arco dado subtiende el mismo ángulo desde todo punto O de aquélla. En la figura 96, el ángulo AOB, subtendido por el arco AB es independiente de la posición de O. Puede relacionarse este hecho con el concepto proyectivo de razón doble, considerando no dos puntos A y B, sino cuatro A, B, C, D de la circunferencia. Las cuatro semirrectas a, b, c, d, que los unen a un quinto punto O de aquélla tendrán una razón doble (a, b, c, d) que dependerá exclusivamente de los ángulos subtendidos por los arcos CA, CB, DA, DB. Si unimos A, B, C, D con otro punto O' de la circunferencia, obtendremos cuatro rayos a', b', c', d'. De acuerdo con la propiedad que acabamos de mencionar, ambas cuaternas de rayos serán «congruentes»[1]; o sea, tendrán la misma razón doble $(a', b', c', d') = (a, b, c, d)$. Si ahora proyectamos la circunferencia en una cónica K, obtendremos cuatro puntos sobre K que llamaremos A, B, C, D, otros dos puntos O y O' y dos cuaternas de rayos a, b, c, d y a', b', c', d'. Éstas no serán congruentes, puesto que, en general, la proyección no conserva la igualdad de los ángulos. Pero, ya que la razón doble permanece invariante en la proyección, la igualdad $(a, b, c, d) = (a', b', c', d')$ seguirá siendo válida. Esto nos conduce al teorema fundamental: *Si se unen cuatro puntos cualesquiera* A, B, C, D *de una cónica* K *con un quinto punto* O *de* K *mediante las rectas* a, b, c, d, *la razón doble* (a, b, c, d) *es independiente de la posición de* O *sobre* K (Figura 97).

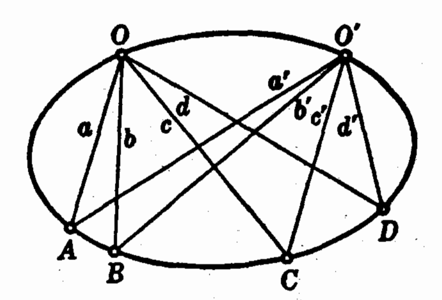


FIG. 97.—Razones dobles en una elipse.

Es éste en verdad un resultado notable; ya sabíamos que cuatro puntos cualesquiera de una recta dan lugar a la misma razón doble desde cualquier quinto punto exterior O. Este teorema sobre razones dobles es el hecho fundamental en geometría proyectiva. Vemos ahora que sigue siendo válido para cuatro puntos de una cónica, con una importante restricción; el quinto punto no puede ser uno cualquiera del plano, aunque puede moverse libremente sobre la cónica dada.

No es difícil probar el teorema recíproco, que aparece en la siguiente

[1] Una cuaterna de rayos, a, b, c, d, se dice congruente con otra a', b', c', d', si el ángulo de cada par de rayos de la primera cuaterna es igual y tiene el mismo sentido que el correspondiente en la segunda cuaterna.

forma: si dos puntos O y O' de una curva K son tales que cualquier cuaterna de puntos A, B, C, D de K tiene la misma razón doble, tanto desde O como desde O', K es una cónica (y, en consecuencia, A, B, C, D tienen la misma razón doble desde un tercer punto cualquiera O'' de K). Omitimos la demostración.

Estas propiedades proyectivas de las cónicas sugieren un método general para construirlas. Llamaremos *haz de rectas* al conjunto de todas las rectas de un mismo plano que pasan por un punto O. Consideremos ahora dos haces de vértices O y O', elegidos sobre la cónica K. Podemos establecer una correspondencia biunívoca entre las rectas del haz O y las del haz O', haciendo corresponder a toda recta a de O, otra a' de O', siempre que a y a' se corten en un punto A de la cónica K. Entonces, cuatro rayos cualesquiera a, b, c, d del haz O tendrán la misma razón doble que los cuatro rayos correspondientes a', b', c', d' de O'. Cualquier correspondencia biunívoca entre dos haces de rectas, que tenga esta propiedad, se llama *correspondencia proyectiva*. (Evidentemente, esta definición es dual de la dada en la pág. 190 para la correspondencia proyectiva entre los puntos de dos rectas.) Se dice que dos haces están relacionados proyectivamente cuando se establece entre ellos una correspondencia proyectiva; con esta definición podemos afirmar ahora que la cónica K es el lugar geométrico de las intersecciones de los rayos correspondientes de dos haces relacionados proyectivamente. Este teorema nos procura la base para una definición puramente proyectiva de las cónicas: *Una cónica es el lugar geométrico de las intersecciones de los rayos correspondientes de dos haces relacionados proyectivamente*[1]. Es tentador seguir el camino hacia la teoría de las cónicas que nos abre esta definición, pero nos limitaremos a unas pocas observaciones.

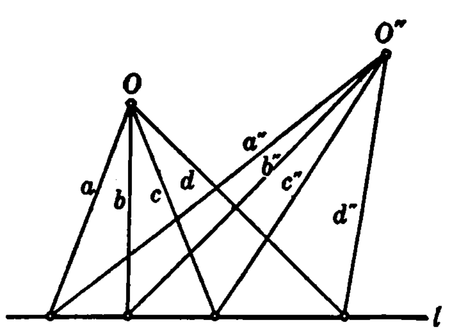


FIG. 98.—Construcción de haces relacionados proyectivamente.

Pueden obtenerse pares de haces, relacionados proyectivamente, de la manera siguiente: proyectemos todos los puntos P de una recta l desde dos centros diferentes O y O''; en los haces proyectantes esta-

[1] Este lugar puede degenerar, en determinadas circunstancias, en una recta (véase figura 98).

blezcamos una correspondencia entre las rectas a y a'' que se cortan sobre l. Ambos haces estarán entonces relacionados proyectivamente. Tomemos ahora el haz O'' y transportémoslo rígidamente a cualquier otra posición O'; el haz resultante O' estará relacionado proyectivamente con el O. Por otra parte, de esta manera puede obtenerse cualquier correspondencia proyectiva entre los haces. (Este hecho es dual del ejercicio 1, pág. 190.) Si los haces O y O' son congruentes, obtenemos una circunferencia. Si los ángulos son iguales, pero de sentido opuesto, la cónica es una hipérbola equilátera (Fig. 99).

Obsérvese que esta definición de cónica puede conducir a un lugar que es una recta, como en la figura 98. En este caso, la recta OO'' se corresponde a sí misma y se admite que todos sus puntos pertenecen

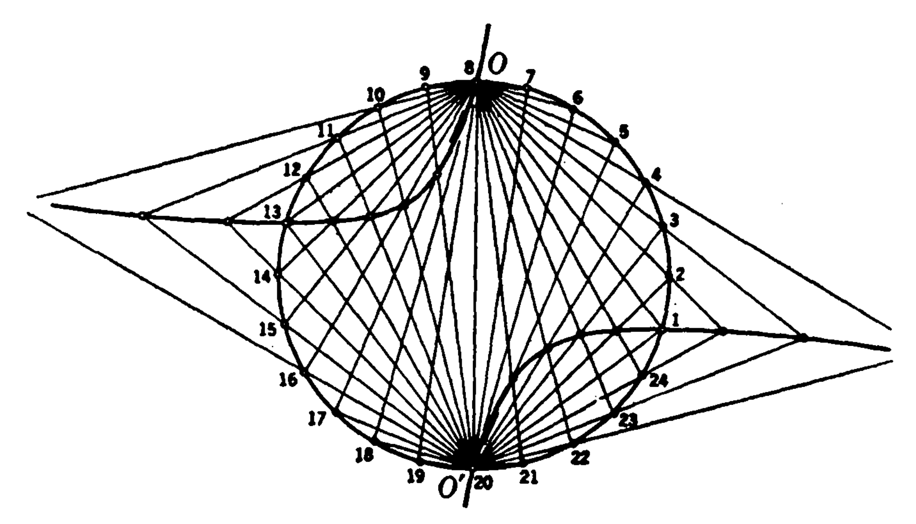


FIG. 99.—Generación de la circunferencia e hipérbola equilátera mediante haces proyectivos.

al lugar. Por tanto, la cónica degenera en un par de rectas, lo que está de acuerdo con la existencia de secciones de un cono (las obtenidas por planos que pasan por el vértice) que se componen de dos rectas.

Ejercicios:

1. Dibújense elipses, hipérbolas y parábolas mediante haces proyectivos. (Se aconseja al lector que se habitúe a estas construcciones, pues contribuirán mucho a una comprensión perfecta del tema.)

2. Dados cinco puntos O, O', A, B, C, de una cónica desconocida K, se desea determinar el punto D en el que una recta dada d, que pasa por O, corta a K. (*Indicación:* Considérense los rayos a, b, c, que pasan por O, dados mediante OA,

OB, OC, y, análogamente, los rayos a', b', c', que pasan por O'. Trácese por O el rayo d y constrúyase por O' el rayo d' tal que $(a, b, c, d) = (a', b', c', d')$. La intersección de d y d' es necesariamente un punto de K.)

3. Las cónicas como envolventes.—El concepto de tangente a una cónica pertenece a la geometría proyectiva, pues se trata de una recta que tiene con la cónica un solo punto común, propiedad que se conserva en la proyección. Las propiedades proyectivas de las tangentes a las cónicas se basan en el siguiente teorema fundamental: *la razón doble de los puntos de intersección de cuatro tangentes fijas a una cónica con una quinta tangente, es la misma para cualquier posición de esta última.*

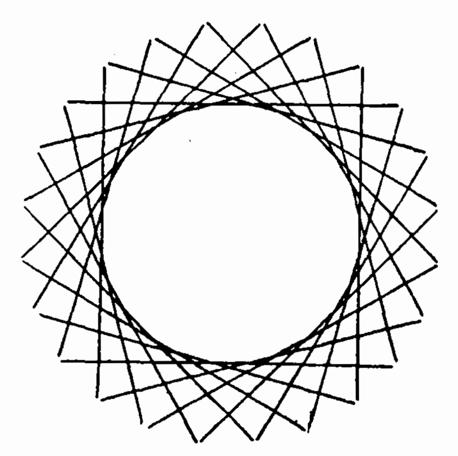


Fig. 100.—La circunferencia, definida por el conjunto de sus tangentes.

La demostración de este teorema es muy sencilla; puesto que una cónica es la proyección de una circunferencia, y dado que el teorema se refiere únicamente a propiedades que permanecen invariantes en la proyección, bastará demostrarlo para la circunferencia, y quedará así establecido en general. Para la circunferencia el teorema es sólo un problema de geometría elemental.

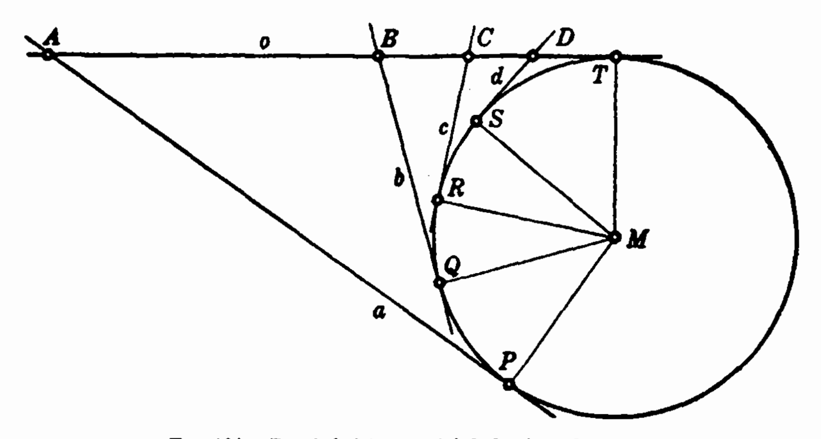


Fig. 101.—Propiedad tangencial de la circunferencia.

Sean P, Q, R, S cuatro puntos cualesquiera de una circunferencia K, siendo a, b, c, d las tangentes correspondientes. Sea T otro punto con tangente o, que corta a a, b, c, d en A, B, C, D. Si M es el centro de la circunferencia, se tendrá evidentemente que $\widehat{TMA} = {}^1/_2\, \widehat{TMP}$, siendo este último igual al ángulo subtendido por el arco TP desde un punto de K. Análogamente, $\widehat{TMB}$ es el ángulo subtendido por el arco TQ desde un punto de K. En consecuencia, $\widehat{AMB} = {}^1/_2\widehat{PQ}$, siendo ${}^1/_2\widehat{PQ}$ el ángulo subtendido por el arco PQ desde un punto de K. De ahí que los puntos A, B, C, D se proyecten desde M en cuatro

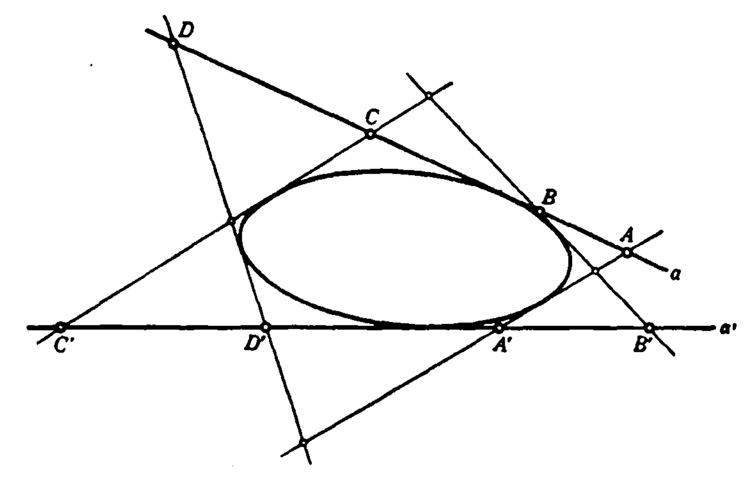


FIG. 102.—Series proyectivas de puntos sobre dos tangentes de una elipse.

rayos, cuyos ángulos están dados por las posiciones fijas de P, Q, R, S. Resulta, por tanto, que la razón doble $(ABCD)$ depende sólo de las cuatro tangentes a, b, c, d y no de la posición particular de la quinta tangente o. Éste es exactamente el teorema que queríamos demostrar.

En la sección precedente hemos visto que podía construirse una cónica mediante los puntos de intersección de las rectas correspondientes de dos haces proyectivos. El teorema que acabamos de demostrar nos permite dar la construcción dual. Sean dos tangentes a y a' de una cónica K. Otra tercera tangente t cortará a a y a' en dos puntos A y A', respectivamente. Si t se mueve sobre la cónica, quedará establecida una correspondencia $A \longleftrightarrow A'$ entre los puntos de a y los de a'; esta correspondencia es proyectiva, pues de acuerdo con nuestro teorema cuatro puntos cualesquiera de a tendrán la misma

razón doble que los correspondientes de *a'*. Resulta de ello que una *cónica K*, considerada como el conjunto de sus tangentes, *se compone de las rectas que unen los pares de puntos correspondientes de las dos*

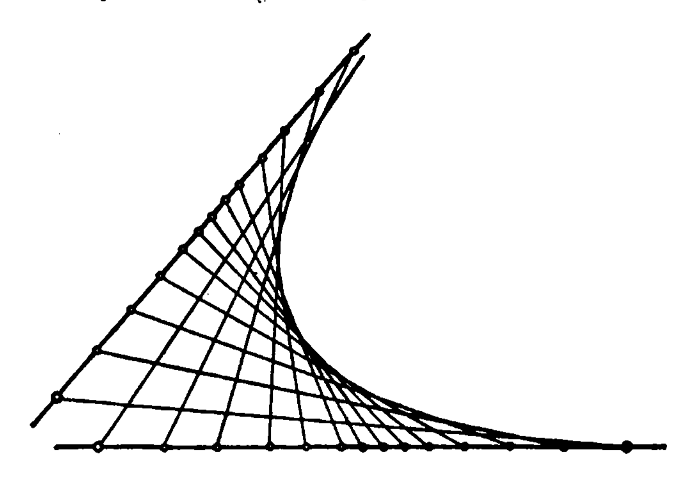


FIG. 103.—La parábola definida por series congruentes de puntos.

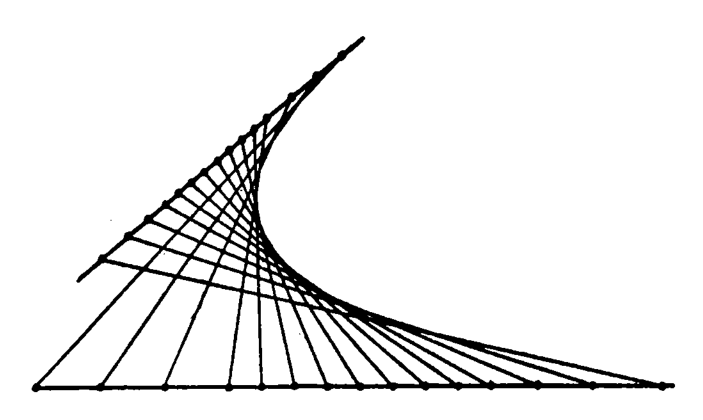


FIG. 104.—La parábola definida por series semejantes de puntos.

series[1] *de puntos proyectivos sobre* a *y* a'. Puede utilizarse esta propiedad para dar una definición proyectiva de las cónicas como «envolventes de rectas». Comparémosla con la definición proyectiva de cónica, dada en la sección precedente:

[1] El conjunto de puntos de una recta se llama *serie de puntos*. Es el concepto dual de haz de rectas.

I. Una cónica es un conjunto de *puntos*, formado por los *puntos de intersección de las rectas* correspondientes de dos *haces* proyectivos de *rayos*.

II. Una cónica es un conjunto de *rectas*, formado por las *rectas que unen los puntos* correspondientes de dos *series* proyectivas de *puntos*.

Si tomamos la tangente a una cónica en un punto cualquiera como el elemento dual del propio punto de contacto, y si consideramos una curva «envolvente» (conjunto de todas sus tangentes) como dual de una curva «puntual» (conjunto de todos sus puntos), es evidente la completa dualidad de ambas proposiciones. En la traducción de un enunciado a otro, reemplazando cada concepto por su dual, la voz «cónica» no varía; en un caso es una «cónica puntual», definida por sus puntos; en el otro, una «cónica envolvente», definida por sus tangentes (véase la Fig. 100).

Una consecuencia importante de este hecho es que el principio de dualidad de la geometría proyectiva plana, enunciado inicialmente sólo para puntos y rectas, puede ahora generalizarse hasta comprender a las cónicas. *Si, en el enunciado de un teorema cualquiera relativo a puntos, rectas y cónicas, se reemplaza cada elemento por su dual* (recordando que el dual de un punto de una cónica es la tangente a la misma) *el resultado será también cierto*. Un ejemplo de aplicación de este principio lo hallaremos en el artículo siguiente.

La construcción de las cónicas como envolventes se ilustra en las figuras 103 y 104. Si en las dos series proyectivas de puntos se corresponden entre sí los dos puntos del infinito (como debe ocurrir en series congruentes[1] o semejantes), la cónica será una parábola; el teorema recíproco es también cierto.

Ejercicio: Demuéstrese el teorema recíproco: sobre dos tangentes fijas cualesquiera de una parábola, una tangente móvil determina dos series de puntos semejantes.

4. Los teoremas generales de Pascal y Brianchon para las cónicas. Uno de los más interesantes ejemplos del principio de dualidad para las cónicas es la relación entre los teoremas generales de Pascal y Brianchon. El primero fué descubierto en 1640, y el segundo en 1806. Sin embargo, el uno es consecuencia inmediata del otro, ya que cualquier teorema que se refiera exclusivamente a cónicas, rectas y puntos sigue verificándose si se sustituye por el enunciado dual.

Los teoremas dados en las páginas 200 y 202 bajo la misma denomi-

[1] Es obvio lo que se quiere decir por correspondencia *congruente* o *semejante* entre dos series de puntos.

nación son casos especiales de los siguientes teoremas más generales:

Teorema de Pascal: Los lados opuestos de un hexágono inscrito en una cónica se encuentran en tres puntos alineados.

Teorema de Brianchon: Las tres diagonales que unen los vértices opuestos de un hexágono circunscrito a una cónica son concurrentes.

Es obvio que ambos teoremas son de carácter proyectivo. Su dualidad se hace evidente en cuanto se formulan como sigue:

Teorema de Pascal: Dados seis puntos 1, 2, 3, 4, 5, 6, de una cónica, se unen los puntos sucesivos mediante las rectas (1, 2), (2, 3),

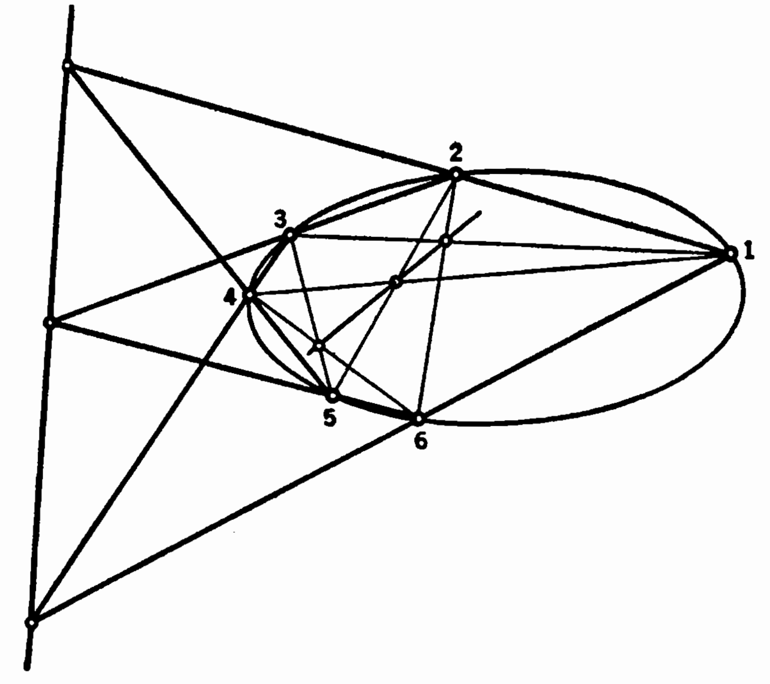


Fig. 105.—Configuración general de Pascal para dos hexágonos (1, 2, 3, 4, 5, 6) y (1, 3, 5, 2, 6, 4).

(3, 4), (4, 5), (5, 6), (6, 1). Si se hallan los puntos de intersección de (1, 2) con (4, 5); de (2, 3) con (5, 6), y de (3, 4) con (6, 1), dichos tres puntos de intersección se encuentran sobre una misma recta.

Teorema de Brianchon: Dadas seis tangentes 1, 2, 3, 4, 5, 6, a una cónica, se hallan los puntos de intersección de las tangentes sucesivas (1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 1). Si se trazan las rectas que

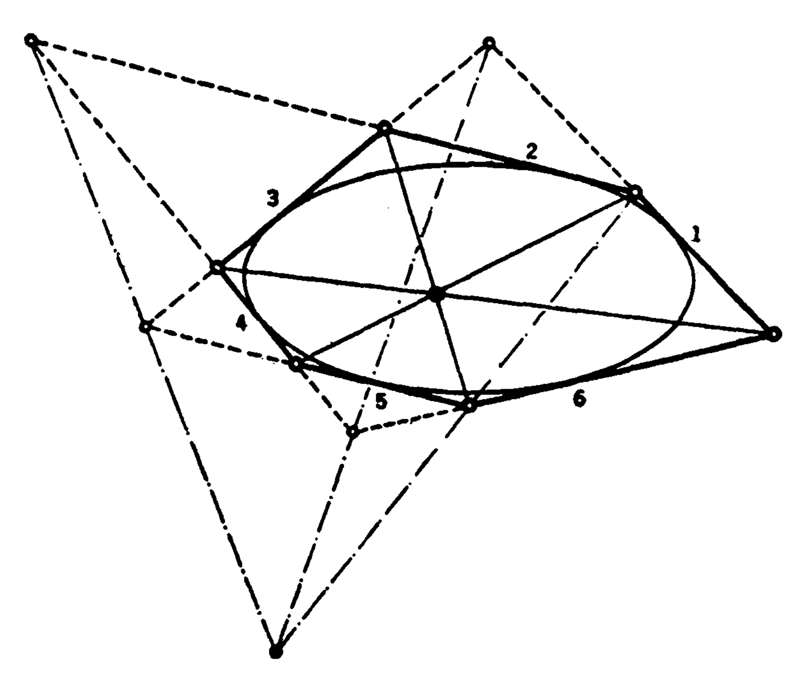


FIG. 106.—Configuración general de Brianchon, también para dos hexágonos.

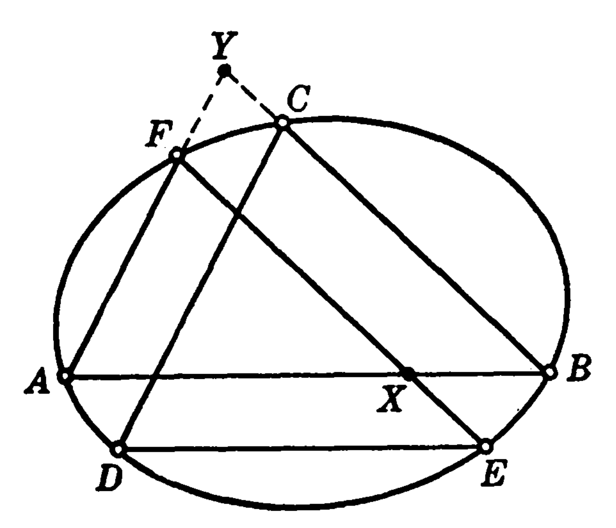


FIG. 107.—Demostración del teorema de Pascal.

unen (1, 2) con (4, 5); (2, 3) con (5, 6), y (3, 4) con (6, 1), dichas tres rectas pasan por un mismo punto.

Las demostraciones resultan de una particularización análoga a la que se utilizó en los casos degenerados. Para demostrar el teorema de Pascal, sean A, B, C, D, E, F los vértices de un hexágono inscrito en una cónica K. Mediante una proyección, podemos hacer que AB sea paralela a ED y FA paralela a CD, con lo que obtenemos la configuración representada en la figura 107. (Para mayor comodidad se ha representado el hexágono como si se cortara a sí mismo, aunque esto no es necesario.) El teorema de Pascal se reduce ahora a la sencilla proposición de que CB es paralela a FE, o dicho de otra manera, que la recta sobre la cual se hallan las intersecciones de los lados opuestos del hexágono es la recta del infinito. Para demostrarlo, consideremos los puntos F, A, B, D, que, como ya sabemos, se proyectan mediante cuatro rayos cuya razón doble es constante, k, desde cualquier otro punto de K, p. ej., desde C o desde E. Proyectemos estos puntos desde C; entonces, los rayos proyectantes cortan a AF en cuatro puntos F, A, Y, ∞, cuya razón doble es k. O sea, que $YF : YA = k$ (véase pág. 197). Si se proyectan las mismos puntos desde E sobre BA, se tiene:

$$k = (XAB\,\infty) = BX : BA.$$

Por tanto,

$$BX : BA = YF : YA,$$

con lo que queda establecido el paralelismo de YB y FX. Esto completa la demostración del teorema de Pascal.

El teorema de Brianchon se deduce, sea mediante el principio de dualidad o por un razonamiento directo dual del anterior. Será un excelente ejercicio para el lector establecer con detalle la demostración.

5. El hiperboloide.—Las figuras que corresponden en tres dimensiones a las cónicas del plano son las «cuádricas», entre las cuales la esfera y el elipsoide son casos especiales. Estas superficies ofrecen mayor variedad y su estudio resulta considerablemente más difícil que el de las cónicas. Aquí estudiaremos brevemente, y sin dar las demostraciones, una de las cuádricas más interesantes: el «hiperboloide de una hoja».

Esta superficie puede definirse de la siguiente forma: elíjanse tres rectas cualesquiera, l_1, l_2, l_3, sin ninguna particularidad en cuanto a su posición en el espacio; es decir, tales que dos cualesquiera de ellas no sean concurrentes o paralelas, ni que las tres sean paralelas a un

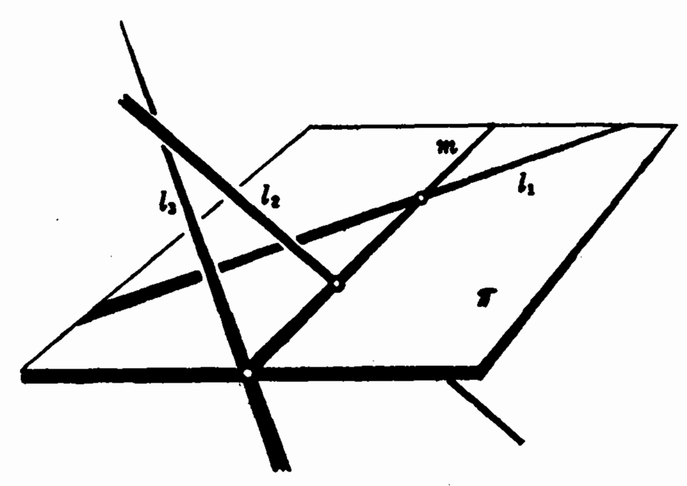


FIG. 108.—Construcción de rectas que se apoyan en tres dadas en posición general.

mismo plano. Es sorprendente que existan infinitas rectas en el espacio que se apoyan en las tres dadas. Para ver esto, tomemos un plano cualquiera π que pase por l_1. Entonces, π cortará a l_2 y l_3 en dos puntos, y la recta m que los une cortará a l_1, l_2 y l_3, engendrando una superficie indefinida, que es el hiperboloide de una hoja, el cual contiene una familia infinita de rectas de la clase m. Cualquier terna de estas rectas m_1, m_2, m_3 estará también en posición general, y todas las rectas del espacio que corten a estas tres se hallarán contenidas en la superficie del hiperboloide. Ésta es la propiedad fundamental de esta superficie: está compuesta de dos familias diferentes de rectas; cada tres de la misma familia están en posición general, mientras cada recta de una familia corta a todas las de la otra.

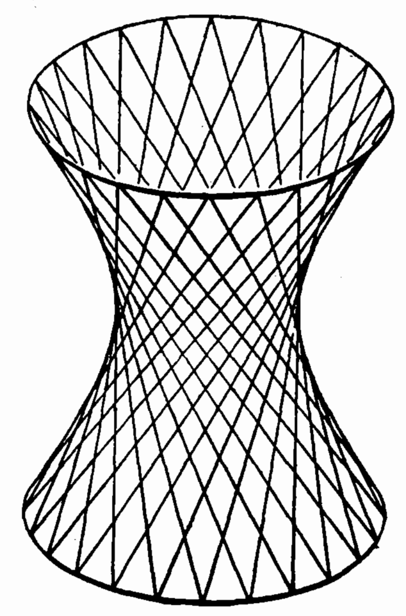


FIG. 109.—El hiperboloide.

Una importante propiedad

proyectiva del hiperboloide consiste en que la razón doble de cuatro puntos, intersecciones de cuatro rectas cualesquiera de una familia con una de la otra, es independiente de la posición de esta última. Esto se deduce directamente del método de construcción del hiperboloide mediante un plano que gira, lo que puede demostrar el lector como ejercicio.

Una de las propiedades más notables del hiperboloide es la de que, si bien contiene dos familias de rectas que se cortan, esto no hace que la superficie sea rígida. Si se construye un modelo con alambres que puedan girar libremente en los puntos de intersección, puede deformarse la figura de manera continua, adquiriendo una gran variedad de formas.

IX. AXIOMÁTICA Y GEOMETRÍA NO EUCLÍDEA

1. El método axiomático.—El método axiomático en matemáticas se remonta por lo menos a Euclides. De ningún modo es cierto que la matemática griega estuviese desarrollada o presentada exclusivamente en la forma rígida de postulados de los *Elementos*. Pero tan grande fué la impresión producida por este trabajo sobre las generaciones subsiguientes, que se transformó en un modelo para toda demostración rigurosa en matemáticas. A veces hasta los filósofos como, por ejemplo, Spinoza en su *Ethica, more geometrico demonstrata*, trató de presentar sus argumentos en forma de teoremas deducidos de definiciones y axiomas. En la matemática moderna, después de un alejamiento de la tradición euclídea durante los siglos XVII y XVIII, hubo una creciente penetración del método axiomático en todos los campos. Uno de los resultados más recientes ha sido la creación de una nueva disciplina: la lógica matemática.

En términos generales, el punto de vista axiomático puede describirse como sigue: Probar un teorema en un sistema deductivo consiste en hacer ver que el teorema es una consecuencia lógica y necesaria de ciertas proposiciones previamente establecidas, que a su vez deben ser probadas, y así sucesivamente. El proceso de demostración matemática sería, por tanto, una tarea imposible de regresión infinita, salvo que, en esta marcha hacia atrás, esté permitido detenerse en algún punto. Por tanto, debe de haber un número de proposiciones, llamadas *postulados* o *axiomas*, que son aceptadas como verdaderas, y para las cuales no se requiere ninguna demostración. De éstas podemos intentar deducir todos los teoremas por medio de argumentos puramente lógicos. Si los hechos de un campo científico pueden colocarse en un orden lógico tal que sea posible deducir todo de un cierto

número de aserciones previamente elegidas (preferiblemente pocas, sencillas y adecuadas), entonces se dice que ese campo está presentado en forma axiomática. La elección de las proposiciones aceptadas como axiomas es en gran medida arbitraria; pero poco se gana con el método axiomático, si los postulados no son simples y no demasiado numerosos. Además, los postulados deben ser *compatibles*, en el sentido de que no puedan deducirse de ellos dos teoremas contradictorios entre sí; y *suficientes*, de tal modo que todo teorema del sistema sea deducible de ellos. Por razones de economía es también deseable que los postulados sean *independientes*, en el sentido de que ninguno de ellos sea una consecuencia lógica de los restantes.

La cuestión de la compatibilidad y de la suficiencia de un conjunto de axiomas ha sido objeto de mucha controversia. Diferentes convicciones filosóficas concernientes a los últimos cimientos del conocimiento humano han llevado a actitudes aparentemente irreconciliables sobre los fundamentos de la matemática. Si los entes matemáticos son considerados como objetos sustanciales en un reino de «pura intuición», independientes de las definiciones y de los actos individuales de la mente humana, entonces, por supuesto, no puede haber contradicciones, ya que los hechos matemáticos son aserciones objetivamente verdaderas, que describen realidades existentes. Desde este punto de vista «kantiano», no hay problema de compatibilidad. Por desgracia, sin embargo, el cuerpo efectivo de la matemática no puede ser adaptado a un sistema filosófico tan simple. Los modernos matemáticos intuicionistas no confían en la intuición pura en el vasto sentido kantiano. Aceptan el infinito numerable como hijo legítimo de la intuición, y admiten sólo propiedades constructivas; pero de este modo, conceptos básicos, como el continuo numérico, serían desterrados; partes importantes de la matemática excluídas, y el resto desesperadamente complicado.

Muy diferente es el punto de vista adoptado por los «formalistas». Éstos no atribuyen una realidad intuitiva a los objetos matemáticos, ni proclaman que los axiomas expresan verdades obvias concernientes a las realidades de la intuición pura; su interés radica sólo en el procedimiento lógico formal del razonamiento sobre la base de los postulados. Esta actitud tiene una definida ventaja sobre el intuicionismo, puesto que garantiza a la matemática toda la libertad necesaria para la teoría y las aplicaciones. Pero ello impone al formalista la necesidad de probar que sus axiomas, que ahora parecen creaciones arbitrarias de la mente humana, no pueden llevarnos a una contradicción. Se han realizado grandes esfuerzos durante los últimos

veinte años para encontrar tales pruebas de compatibilidad, al menos para los axiomas de la aritmética y del álgebra, y para el concepto del continuo numérico. Los resultados son altamente significativos, pero el triunfo está todavía lejos. Realmente, los resultados recientes indican qué tales esfuerzos no pueden ser completamente satisfactorios, en el sentido de que las pruebas de compatibilidad y suficiencia no son posibles dentro de sistemas estrictamente cerrados de conceptos. Es digno de observarse que en todos estos argumentos acerca de los fundamentos se procede por métodos que son enteramente constructivos en sí mismos y dirigidos por esquemas intuitivos.

Acentuado por las paradojas de la teoría de conjuntos (cap. II, pág. 96), el conflicto entre intuicionistas y formalistas obtuvo gran publicidad por obra de apasionados partidarios de ambas escuelas. El mundo matemático fué sacudido por el grito de «crisis en los fundamentos»; pero la alarma no era, y no debe ser, tomada demasiado en serio. Aun reconociendo todo el mérito debido a los resultados obtenidos en la lucha por la depuración de los fundamentos, sería completamente injustificado inferir que el cuerpo vivo de la matemática haya sido siquiera amenazado por tales diferencias de opinión o por paradojas inherentes a una incontrolada tendencia hacia la generalización ilimitada.

Aparte de las consideraciones filosóficas y del interés en los fundamentos, el método axiomático en matemática resulta el camino natural para desenredar la madeja de interconexiones entre los distintos hechos, y para demostrar el esqueleto lógico esencial de la estructura. Sucede a veces que tal concentración en la estructura formal, con preferencia al significado intuitivo de los conceptos, hace más fácil dar con generalizaciones y aplicaciones que podrían haber pasado inadvertidas de haber utilizado un método más intuitivo. Pero rara vez se obtiene un descubrimiento significativo o una visión esclarecedora si se hace uso de un procedimiento exclusivamente axiomático. El pensamiento constructivo, guiado por la intuición, es la verdadera fuente de la dinámica matemática. A pesar de que la forma axiomática es un ideal, es una peligrosa falacia creer que la axiomática constituye *la* esencia de la matemática. La intuición constructiva de los matemáticos da a esta ciencia un elemento no deductivo e irracional que la hace comparable con la música y el arte.

Desde los días de Euclides, la geometría ha sido el prototipo de una disciplina axiomatizada. Durante siglos, el sistema de axiomas de Euclides ha sido objeto de estudio intensivo. Pero sólo recientemente se ha puesto de manifiesto que sus postulados debían ser modi-

ficados y completados si se deseaba que toda la geometría elemental fuera deducible de ellos. En las postrimerías del siglo XIX, por ejemplo, Pasch descubrió que el orden de los puntos sobre una recta, la noción de «entre», requiere un postulado especial. Pasch formuló la siguiente proposición como nuevo axioma: Una recta que corta un lado de un triángulo en cualquier punto distinto de un vértice debe cortar también a otro lado del triángulo. (Pasar por alto tales detalles lleva a muchas aparentes paradojas, en las que consecuencias absurdas —p. ej., la conocida «demostración» de que todo triángulo es isósceles— parecen deducirse rigurosamente de los axiomas de Euclides. Esto se debe, por lo general, a figuras mal dibujadas, cuyas líneas parecen interceptar interior o exteriormente a ciertos triángulos o circunferencias, cuando en realidad no lo hacen.)

En su famoso libro *Grundlagen der Geometrie*[1] (la primera edición fué publicada en 1901), Hilbert dió un conjunto satisfactorio de axiomas para la geometría, y al propio tiempo hizo un estudio exhaustivo de su mutua independencia, compatibilidad y suficiencia.

En todo sistema de axiomas deben entrar ciertos conceptos no definidos, como *punto* y *recta* en geometría. Su «significado» o relación con objetos del mundo físico no es *matemáticamente* esencial. Pueden ser considerados como entes puramente abstractos, cuyas propiedades matemáticas en un sistema deductivo están dadas completamente por las relaciones existentes entre ellos, enunciadas por los axiomas; p. ej., en geometría proyectiva podemos comenzar con los conceptos no definidos de *punto*, *recta* e *incidencia* o *pertenencia* y con los dos axiomas duales: «Dos puntos distintos cualesquiera pertenecen a una sola recta» y «dos rectas distintas cualesquiera inciden en un solo punto». Desde el punto de vista de la axiomática, la forma dual de tales axiomas es la verdadera fuente del principio de dualidad de la geometría proyectiva. Todo teorema que contenga en su enunciado y en su demostración sólo elementos relacionados por axiomas duales, debe admitir otro teorema dual. En efecto, la demostración del teorema original consiste en la aplicación sucesiva de ciertos axiomas, y la aplicación de los axiomas duales en el mismo orden debe darnos la demostración del teorema dual.

La totalidad de los axiomas de la geometría nos da la *definición implícita* de todos los términos geométricos «no definidos», como «recta», «punto», «incidencia», etc. Para las aplicaciones, es importante el hecho de que los conceptos y los axiomas de la geometría se correspon-

---

[1] Existe edición española: *Fundamentos de la Geometría*, trad. de F. Cebrián. Instituto Jorge Juan de Matemáticas. C.S.I.C., Madrid.

dan con aserciones físicamente verificables acerca de objetos «reales», tangibles. La realidad física a que alude el concepto de «punto» es la de cualquier objeto muy pequeño, como la huella de un lápiz; mientras una «recta» es una abstracción hecha a partir de un hilo tenso o un rayo de luz. La experiencia nos dice que las propiedades de estos puntos y rectas físicas concuerdan *grosso modo* con los axiomas formales de la geometría. Cabe concebir que experimentos mucho más precisos puedan plantear la necesidad de modificar esos axiomas si éstos han de ser adecuados para describir los fenómenos físicos. Pero si los axiomas formales no concuerdan aproximadamente con las propiedades de los objetos físicos, entonces la geometría apenas tendría interés. Así, incluso para los formalistas, hay una autoridad aparte de la mente humana que decide la dirección del pensamiento matemático.

2. Geometría no euclídea hiperbólica.—Hay un axioma de la geometría euclídea cuya «verdad», o sea, cuya correspondencia con datos empíricos acerca de hilos tensos o rayos de luz, no es de manera alguna evidente. Se trata del famoso *postulado de la paralela única*, el cual establece que por un punto exterior a una recta dada se puede trazar *una, y sólo una*, paralela a la misma. El rasgo característico de este postulado consiste en que hace una aserción acerca de *toda* la extensión de una recta, imaginada como indefinida en ambas direcciones; ya que decir que dos rectas son paralelas equivale a afirmar que no se cortan nunca, por mucho que se las prolongue. No está de más decir que hay muchas rectas que pasan por un punto y no cortan a otra recta dada *dentro de una distancia finita dada*, por grande que sea. Dado que la máxima longitud posible de una regla real, de un hilo, o incluso de un rayo de luz visible con un telescopio es ciertamente finita, y puesto que dentro de cualquier círculo finito hay infinitas rectas que pasan por un punto dado y que no cortan a otra recta dada interior al círculo, se deduce que este axioma no podrá ser verificado por experimentación. Todos los otros axiomas de la geometría euclídea tienen carácter finito en el sentido de que tratan con porciones finitas de rectas y con figuras planas de extensión finita. El hecho de que el axioma de las paralelas no sea verificable experimentalmente plantea la cuestión de ver si es o no *independiente* de los otros axiomas. Si fuera una consecuencia lógica necesaria de los otros, sería posible eliminarlo como axioma y dar una demostración del mismo mediante los otros axiomas de Euclides. Durante varios siglos, los matemáticos trataron de hallar esa demostración, porque existía el sentimiento general de que el postulado de las paralelas era de carácter esencialmente diferente de los demás, al faltarle ese carácter de evidente

plausibilidad que debería poseer todo axioma de la geometría. Uno de los primeros intentos de esta naturaleza fué hecho por Proclo (siglo IV a. de J.C.), un comentador de Euclides, quien trató de descartar la necesidad de un postulado especial, relativo a las rectas paralelas, *definiendo* la recta paralela a otra como el lugar de los puntos que se encuentran a una distancia fija de la dada. Proclo no vió que la dificultad se había desplazado a otro lugar, porque sería ahora necesario demostrar que el lugar de tales puntos es en efecto una recta. Como Proclo no pudo probar esto, debió de aceptarlo como postulado en lugar del axioma de las paralelas, y nada se ganó con ello, pues no es difícil demostrar que ambos son equivalentes. El jesuíta Saccheri (1667-1733), y más tarde Lambert (1728-1777), trataron de probar el postulado de las paralelas por el método indirecto de admitir lo contrario y deducir consecuencias absurdas. Lejos de ser absurdas, sus conclusiones realmente equivalían a teoremas de la geometría no euclídea desarrollada después. Si hubieran considerado tales conclusiones, no como absurdas, sino más bien como enunciados compatibles en sí mismos, habrían sido los descubridores de la geometría no euclídea.

En aquel tiempo, todo sistema geométrico que no estuviera en absoluto acuerdo con el de Euclides debía considerarse como un evidente desatino. Kant, el filósofo que mayor influjo ejerció en dicho período, formuló esa actitud en la afirmación de que los axiomas euclídeos son inherentes a la mente humana, y, por tanto, tienen una validez objetiva para el espacio «real». Esta creencia en los axiomas de la geometría euclídea como verdades inalterables, existentes en el reino de la intuición pura, fué uno de los dogmas básicos de la filosofía kantiana. Pero a la larga, ni viejos hábitos del pensamiento ni la autoridad filosófica podían reprimir la convicción de que la interminable serie de fracasos en la búsqueda de una demostración del postulado de las paralelas no era debida a una falta de ingenio, sino más bien al hecho de que tal postulado es realmente *independiente* de los otros. (Análogamente, el fracaso en la resolución de la ecuación general de quinto grado por medio de radicales llevó a la sospecha, más tarde verificada, de que tal solución es imposible.) El húngaro Bolyai (1802-1860) y el ruso Lobachevsky (1793-1856) resolvieron la cuestión construyendo con todo detalle una geometría en la cual el postulado de las paralelas no se verifica. Cuando el joven y entusiasta genio Bolyai sometió su memoria a Gauss, «príncipe de los matemáticos», para el reconocimiento tan ansiosamente esperado, fué informado de que Gauss mismo se le había anticipado, pero no había querido publicar sus resultados por temor a una ruidosa publicidad.

¿Qué significa la independencia del postulado de las paralelas? Sencillamente, que es posible construir un sistema compatible de proposiciones «geométricas» referentes a puntos, rectas, etc., deduciéndolas de un conjunto de axiomas en el cual el postulado de las paralelas se haya reemplazado por un postulado contrario. Tal sistema se llama una geometría no euclídea. Fué precisa la fuerza intelectual de Gauss, Bolyai y Lobachevsky para darse cuenta de que tal geometría, basada en un sistema de axiomas no euclidiano, podía ser perfectamente compatible.

Para probar la compatibilidad de la nueva geometría no basta construir un vasto cuerpo de teoremas no euclidianos, como hicieron Bolyai y Lobachevsky. En lugar de ello, hemos aprendido a construir «modelos» de tal geometría que satisfacén todos los axiomas de Euclides, excepto el de las paralelas. El más sencillo de dichos modelos fué dado por Félix Klein, cuyo trabajo en este campo fué estimulado por las ideas del geómetra inglés Cayley (1821-1895). En este modelo pueden trazarse infinitas *rectas paralelas* a una recta dada por un punto exterior; tal geometría es llamada geometría Bolyai-Lobachevsky o *hiperbólica.* (La razón de este último nombre se encontrará en la pág. 238.)

El modelo de Klein se construye considerando primero objetos de la geometría euclídea ordinaria y *bautizando* a algunos de estos objetos y a las relaciones entre ellos de tal modo que resulte una geometría no euclídea. Este modelo debe ser, *eo ipso,* tan carente de contradicción como el sistema euclídeo original, pues se nos ofrece, visto desde otro punto y descrito con otras palabras, igual que el sistema de proposiciones de la geometría euclídea ordinaria. Dicho modelo puede ser fácilmente comprendido por medio de algunos conceptos de geometría proyectiva.

Si sometemos el plano a una transformación proyectiva sobre otro plano, o mejor sobre sí mismo (por posterior coincidencia del punto imagen con el plano original), en general, un círculo y su interior se transformarán en una cónica. Pero es fácil ver (omitimos la demostración) que existen infinitas transformaciones proyectivas del plano en sí mismo tales que un círculo dado y su interior se transformen en sí mismos. Por tales transformaciones, puntos del interior o del contorno son en general llevados a otras posiciones, pero quedan interiores o en el contorno del círculo (p. ej., podemos transformar el centro del círculo en cualquier otro punto interior). Consideremos la totalidad de tales transformaciones. Ciertamente, éstas no dejarán invariantes las formas de las figuras, y por tanto no son desplazamientos rígidos

en el sentido usual; pero ahora damos el paso decisivo al *llamar* a dichas transformaciones «desplazamientos no euclídeos» de la geometría que estamos construyendo. Por medio de estos «desplazamientos» podemos definir la congruencia: dos figuras se *dirán* congruentes si existe un desplazamiento no euclídeo que transforme una en la otra.

El modelo de Klein de la geometría hiperbólica es entonces el siguiente: el «plano» consiste sólo en los puntos interiores a un círculo; los puntos exteriores no se consideran. Cada punto interior del círculo se *llama* «punto» no euclídeo; cada cuerda del círculo se *llama* «recta» no euclídea; los «desplazamientos» y «congruencias» se han definido anteriormente; unir «puntos» y hallar la intersección de «rec-

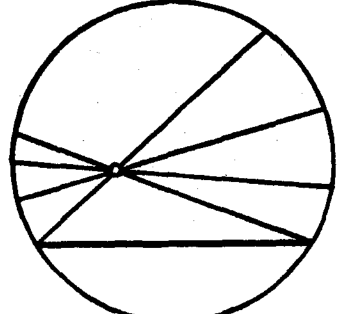


FIG. 110.—Modelo no euclídeo de Klein.

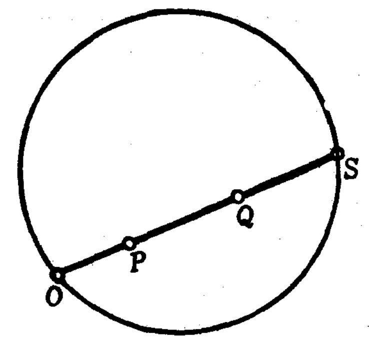


FIG. 111.—Distancia no euclídea.

tas» en el sentido no euclídeo sigue siendo lo mismo que en la geometría de Euclides. Es fácil probar que el nuevo sistema satisface todos los postulados de la geometría euclídea, con la sola excepción del de las paralelas. Que el postulado de las paralelas no se verifica en el nuevo sistema se ve por el hecho de que por todo «punto» exterior a una «recta» pueden trazarse infinitas «rectas» que no tengan ningún «punto» común con la «recta» dada. La primera «recta» es una cuerda euclídea del círculo, mientras la segunda «recta» puede ser cualquiera de las cuerdas que pasan por el «punto» dado y que no cortan a la primera «recta» dentro del círculo. Este sencillo modelo es por completo suficiente para decidir la cuestión fundamental que dió nacimiento a la geometría no euclídea; demuestra que el postulado de las paralelas no puede ser deducido de los otros axiomas de la geometría euclídea; pues si pudiera deducirse, debería ser un teorema verdadero en la geometría del modelo de Klein, y hemos visto que no lo es.

En términos estrictos, este razonamiento está basado en la hipótesis de que la geometría del modelo de Klein no es contradictoria, en el sentido de que no

pueden demostrarse a la vez un teorema y su contrario. Pero esta geometría está, efectivamente, tan desprovista de contradicción como la geometría euclídea ordinaria, ya que las proposiciones referentes a *puntos*, *rectas*, etc., en el modelo de Klein, son meramente modos distintos de enunciar determinados teoremas de la geometría euclídea. No ha sido dada hasta ahora una demostración satisfactoria de la compatibilidad de los axiomas de la geometría euclídea, salvo por reducción a los conceptos de la geometría analítica y por ello, en última instancia, al continuo numérico, cuya carencia de contradicción es de nuevo una cuestión sin resolver.

*Debemos mencionar aquí un detalle que excede a nuestro objetivo inmediato, esto es, la manera de definir la «distancia» no euclídea en el modelo de Klein. Esta «distancia» debe resultar invariante respecto a cualquier «desplazamiento» no euclídeo, pues éste no debe alterar las distancias. Sabemos que las razones dobles no varían en la proyección, y una razón doble que haga intervenir dos puntos arbitrarios P y Q, interiores al círculo, aparece inmediatamente si se prolonga el segmento PQ hasta su intersección con la circunferencia en O y S. La razón doble $(OSPQ)$ de estos cuatro puntos es un número (positivo) que puede tomarse como definición de «distancia» PQ entre P y Q, si bien debemos modificar ligeramente esta definición para hacerla útil. Pues si los tres puntos P, Q, R están sobre una recta, debe verificarse que $\overline{PQ} + \overline{QR} = \overline{PR}$. Ahora bien: en general,

$$(OSPQ) + (OSRQ) \neq (OSRP).$$

Por el contrario, tenemos la relación

$$(OSPQ)\,(OSRQ) = (OSRP), \qquad [1]$$

como resulta de las igualdades

$$(OSQP)\,(OSRQ) = \frac{QO/QS}{PO/PS} \cdot \frac{RO/RS}{QO/QS} = \frac{RO/RS}{PO/PS} = (OSRP).$$

Como consecuencia de la ecuación [1] podemos dar una definición aditiva satisfactoria midiendo la «distancia», no por la razón doble, sino por el *logaritmo de la razón doble*:

$$\overline{PQ} = \text{distancia no euclídea de } P \text{ a } Q = \log\,(OSQP).$$

Esta distancia será un número positivo, ya que $(OSQP) > 1$ si $P \neq Q$. Utilizando la propiedad fundamental del logaritmo (véase Cap. VII) resulta de [1] que $\overline{PQ} + \overline{QR} = \overline{PR}$. Es indiferente la base que se elija para el sistema de logaritmos, pues al variar ésta, solamente cambia la unidad de medida. Observemos que si uno de los puntos, p. ej., Q, se aproxima a la circunferencia, la distancia no euclídea PQ tiende a infinito, lo cual prueba que la recta de nuestra geometría no euclídea

tiene longitud no euclídea infinita, aunque en el sentido euclídeo ordinario sea solamente un segmento finito de recta.

3. Geometría y realidad.—El modelo de Klein demuestra que la geometría hiperbólica, considerada como un sistema formal deductivo, resulta tan satisfactoria como la geometría euclídea clásica. La cuestión que se presenta es: ¿cuál de las dos es preferible como descripción de la geometría del mundo físico? Como ya hemos visto, la experimentación nunca podrá decidir si hay una sola paralela o bien hay infinitas por un punto exterior a una recta dada. En la geometría euclídea, sin embargo, la suma de los ángulos de cualquier triángulo es 180°. A este respecto, Gauss ideó un experimento para resolver la cuestión. Midió con todo cuidado los ángulos de un triángulo formado por tres cimas montañosas muy distantes, y encontró que la suma era de 180°, dentro de los límites del error experimental. Si el resultado hubiera sido notablemente inferior a 180°, la consecuencia inmediata sería que la geometría hiperbólica es preferible para describir la realidad física. Pero nada se resolvió con este experimento, ya que para pequeños triángulos, cuyos lados tienen sólo unos pocos kilómetros de longitud, la diferencia con 180° en la geometría hiperbólica debe ser tan pequeña como para pasar inadvertida con los instrumentos usados por Gauss. Así, a pesar de que el experimento no resolvió la cuestión, probó, no obstante, que las geometrías hiperbólica y euclídea, cuyas diferencias *en grande* son notables, coinciden casi por completo para figuras relativamente pequeñas, por lo que experimentalmente son equivalentes. Por tanto, mientras se consideren propiedades puramente *locales* del espacio, la elección entre ambas geometrías debe hacerse sólo sobre la base de la sencillez y comodidad. Puesto que el sistema euclídeo es más sencillo de manejar, está justificado que lo usemos exclusivamente, por lo menos mientras se consideran distancias pequeñas (¡de unos cuantos millones de kilómetros!). Pero no debemos esperar necesariamente que dicho sistema sea apropiado para describir el universo como un todo, en sus aspectos más grandiosos. Aquí la situación es precisamente análoga a la que existe en física, donde los sistemas de Newton y de Einstein dan los mismos resultados para pequeñas distancias y velocidades, pero divergen cuando se trata de grandes magnitudes.

La importancia revolucionaria del descubrimiento de la geometría no euclídea estriba en el hecho de que destruyó la noción que se tenía de los axiomas de Euclides, como esquema matemático inmutable al que debía adaptarse nuestro conocimiento experimental de la realidad física.

4. Modelo de Poincaré.—El matemático es libre de considerar una «geometría» como definida por cualquier sistema de axiomas compatibles, referentes a *puntos*, *rectas*, etc.; sus investigaciones resultarán útiles al físico únicamente cuando estos axiomas se ajusten al comportamiento físico de los objetos del mundo real. Desde este punto de vista, vamos a examinar el significado de la frase «la luz se propaga en línea recta». Si se considera esto como *definición física* de «recta», entonces los axiomas de la geometría deben ser elegidos de forma que se correspondan con el comportamiento de los rayos de luz. Imaginemos, con Poincaré, un universo constituído por el interior de una circunferencia C y tal que la velocidad de la luz en un punto del círculo sea igual a la distancia del mismo a la circunferencia. Puede demostrarse que los rayos de luz adoptarán la forma de arcos circulares, normales en sus extremos a la circunferencia C. En tal universo, las propiedades geométricas de las «rectas» (definidas como rayos de luz) diferirán de las propiedades euclídeas de las rectas, y, en particular, el postulado de las paralelas dejará de cumplirse, puesto que existen infinitas «rectas» que pasan por un punto y no cortan a una «recta» dada. Por supuesto, que los «puntos» y «rectas» de este universo tendrían exactamente las propiedades geométricas de los «puntos» y «rectas» del modelo de Klein. En otras palabras: tendríamos un modelo diferente de la geometría hiperbólica. Pero la geometría euclídea se aplicará también a este universo; en lugar de ser «rectas» no euclídeas, los rayos luminosos serían circunferencias euclídeas ortogonales a C. Vemos así que diferentes sistemas de geometría pueden describir la misma situación física, con tal que los objetos físicos (en este caso, rayos de luz) estén correlacionados con los diferentes conceptos de los dos sistemas:

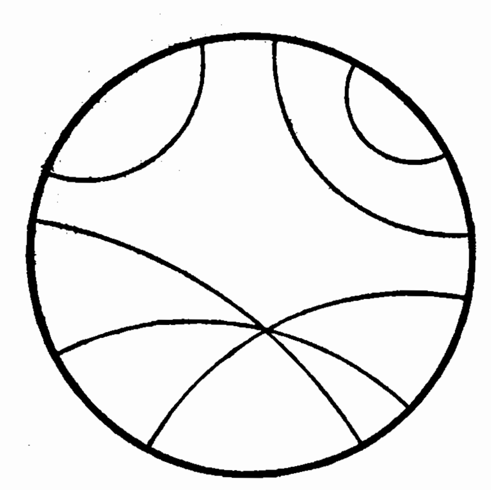


FIG. 112.—Modelo no euclídeo de Poincaré.

rayo de luz → «línea recta» — geometría hiperbólica.
rayo de luz → «circunferencia» — geometría euclídea.

Puesto que el concepto de recta en la geometría euclídea corresponde a la conducta de un rayo de luz en un medio homogéneo, diremos que la geometría de la región interior a C es hiperbólica, para significar sólo que las propiedades físicas de los rayos luminosos en ese universo corresponden a las propiedades de las «rectas» de la geometría hiperbólica.

5. Geometría elíptica o de Riemann.—En la geometría de Euclides, así como en la hiperbólica de Bolyai y Lobachevsky, se supone tácitamente que la recta es infinita (la extensión infinita de la recta está esencialmente vinculada con el concepto y los axiomas de «estar entre»). Pero después que la geometría hiperbólica hubo abierto el camino hacia la libre construcción de geometrías, era natural preguntar si podía construirse una geometría no euclídea en la cual una línea recta no fuera infinita, sino finita y cerrada. Por supuesto, en tales geometrías no sólo debería abandonarse el postulado de las paralelas, sino también los axiomas referentes a «estar entre». Los desarrollos modernos han probado la importancia física de estas geometrías. Fueron consideradas por primera vez en el discurso inaugural pronunciado en 1851 por Riemann con motivo de su admisión como profesor adjunto («Privat-Docent») en la Universidad de Gotinga. Pueden construirse geometrías con rectas finitas y cerradas en una forma carente por completo de contradicción.

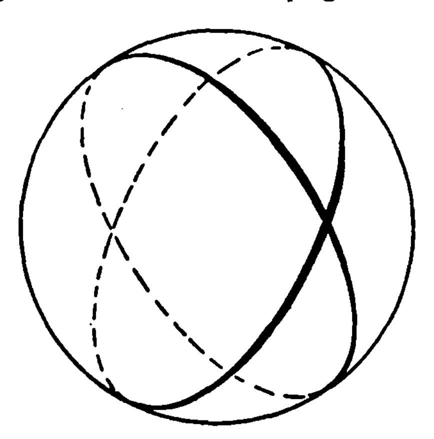


FIG. 113.—«Líneas rectas» en la geometría de Riemann.

Imaginemos un mundo bidimensional, consistente en la superficie S de una esfera, en el cual definimos la *recta* con el significado de círculo máximo de la esfera. Éste sería el camino natural para describir el mundo de un navegante, puesto que los arcos de círculo máximo son las curvas de longitud mínima entre dos puntos de una superficie esférica, y ésta es la propiedad característica de las rectas del plano. En tal mundo, dos «rectas» *cualesquiera* se cortan, de modo que por un punto exterior a una recta no puede trazarse *ninguna* paralela (esto es, no secante) a la «recta» dada. La geometría de las «rectas» en dicho mundo se llama *geometría elíptica*. En esta geometría, la distancia entre dos puntos se mide simplemente por su distancia a

lo largo del menor arco del círculo máximo que los une, mientras los ángulos se miden como en la geometría euclídea. Consideramos en general como típico de una geometría elíptica el hecho de que no exista ninguna paralela a una recta dada.

Siguiendo a Riemann, podemos generalizar dicha geometría como sigue. Consideremos un universo consistente en una superficie curva en el espacio, no necesariamente esférica, y definamos la «recta» que

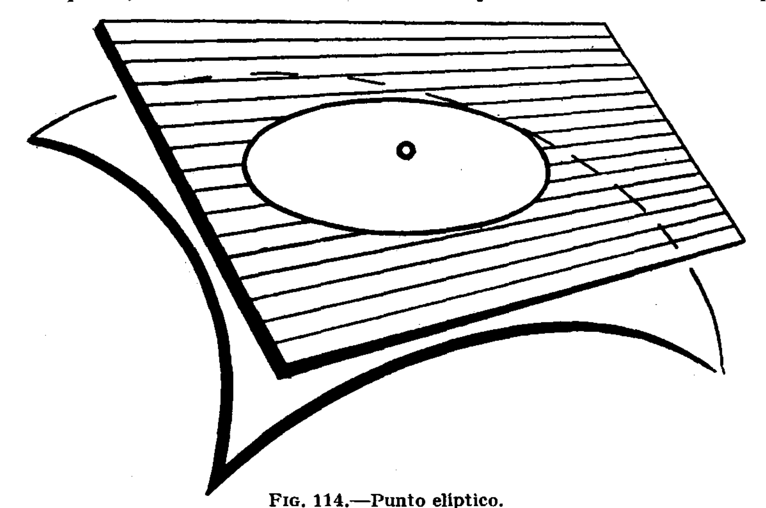


FIG. 114.—Punto elíptico.

une dos puntos como la *curva de menor longitud* o «geodésica» que une esos puntos. Los puntos de la superficie pueden dividirse en dos clases: 1) puntos tales que en un entorno de cada uno la superficie es como una esfera, en el sentido de que está situada por completo de un mismo lado del plano tangente en ese punto; 2) puntos tales que en un entorno de cada uno la superficie tiene forma de silla de montar y está situada a ambos lados del plano tangente en el punto. Los puntos de la primera clase se llaman puntos elípticos de la superficie, puesto que si el plano tangente se traslada un poco paralelamente a sí mismo, corta a la superficie en una curva elíptica; mientras que los puntos de la segunda clase se llaman hiperbólicos, ya que, si el plano tangente se traslada un poco paralelamente a sí mismo, corta a la superficie en una curva parecida a una hipérbola. La geometría de las «rectas» geodésicas en el entorno de un punto de una superficie es elíptica o hiperbólica según que el punto sea elíptico o hiperbólico. En tal modelo de geometría no euclídea, los ángulos se miden por su valor euclídeo ordinario.

Esta idea fué desarrollada por Riemann, quien consideró una geometría del espacio análoga a esta geometría de una superficie, en la cual la «curvatura» del espacio puede cambiar el carácter de la geometría de un punto a otro. Las *rectas* en una geometría de Riemann son las geodésicas. En la teoría general de la relatividad de Einstein, la geometría del espacio es una geometría de Riemann; la luz se propaga a lo largo de geodésicas, y la curvatura del espacio se determina por la naturaleza de la materia que lo llena.

Desde su origen en el estudio de la axiomática, la geometría no euclídea se ha ido desarrollando hasta convertirse en un instrumento útil para su aplicación al mundo físico. En la teoría de la relatividad, en óptica, y en la teoría general de la propagación de ondas, una des-

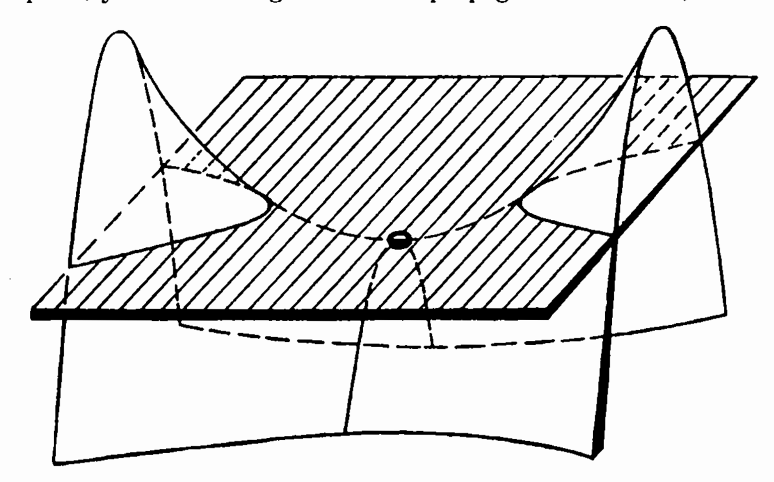


FIG. 115.—Punto hiperbólico.

cripción no euclídea de los fenómenos es a veces mucho más adecuada que la descripción clásica.

APÉNDICE

*GEOMETRÍA DE MÁS DE TRES DIMENSIONES

1. Introducción.—El *espacio real*, que es el medio donde tiene lugar nuestra experiencia física, tiene tres dimensiones; el plano, dos, y la recta, una. Nuestra intuición espacial, en un sentido ordinario, está definitivamente limitada a tres dimensiones. Sin embargo, en muchas ocasiones es conveniente hablar de *espacios* de cuatro o más

dimensiones. ¿Cuál es el significado de un espacio n-dimensional cuando n es mayor que tres, y para qué fines puede servir? Puede darse una respuesta tanto desde el punto de vista analítico como desde el puramente geométrico. La terminología de un espacio n-dimensional puede considerarse meramente como un lenguaje geométrico sugestivo para ideas matemáticas que no están ya al alcance de la intuición geométrica ordinaria. Daremos una breve indicación de las sencillas consideraciones que motivan y justifican este lenguaje.

2. Método analítico.—Hemos insistido ya en la inversión de significado que tuvo lugar en el curso del desarrollo de la geometría analítica. Puntos, rectas, curvas, etc., eran originariamente considerados como entes puramente *geométricos*, y el objeto de la geometría analítica era sólo el de asignarles sistemas de números o ecuaciones, e interpretar o desarrollar una teoría geométrica por métodos algebraicos o analíticos. En el transcurso del tiempo, el punto de vista opuesto comenzó a imponerse cada vez más. Un número x o un par de números x, y o una terna de números x, y, z, fueron considerados como los objetos fundamentales, y estas entidades analíticas fueron luego «interpretadas» como puntos de una recta, de un plano, o del espacio. Desde este punto de vista, el lenguaje geométrico sirve únicamente para establecer relaciones entre números. Podemos descartar el carácter primario o incluso independiente de los objetos geométricos, diciendo que un par de números x, y *es* un punto del plano, que el conjunto de todos los pares de números x, y que satisfacen a la ecuación lineal $L(x, y) = ax + by + c = 0$ de coeficientes fijos a, b, c, *es* una recta, etc. Definiciones análogas pueden darse para el espacio de tres dimensiones.

Aun estando primordialmente interesados en un problema algebraico puede ocurrir que el lenguaje geométrico proporcione una descripción breve y adecuada del mismo, y que la intuición geométrica sugiera el procedimiento algebraico apropiado; p. ej., si deseamos resolver un sistema de tres ecuaciones lineales con tres incógnitas x, y, z:

$$\begin{aligned} L(x, y, z) &= ax + by + cz + d = 0 \\ L'(x, y, z) &= a'x + b'y + c'z + d' = 0 \\ L''(x, y, z) &= a''x + b''y + c''z + d'' = 0, \end{aligned}$$

podemos interpretar el problema como si se tratara de hallar el punto de intersección en el espacio tridimensional R_3 de tres planos definidos por las ecuaciones $L = 0$, $L' = 0$, $L'' = 0$. También, si estamos sólo considerando los pares de números x, y para los cuales $x > 0$, podemos interpretarlos como representando el semiplano de la dere-

cha del eje x. Con mayor generalidad, la totalidad de los pares de números x, y para los cuales

$$L(x, y) = ax + by + d > 0$$

puede interpretarse como el semiplano a un lado de la recta $L = 0$, y la totalidad de las ternas de números x, y, z para las cuales

$$L(x, y, z) = ax + by + cz + d > 0$$

puede considerarse como el «semiespacio» de un lado del plano $L(x, y, z) = 0$.

La introducción de un «espacio tetradimensional», o incluso de un «espacio n-dimensional», resulta ahora natural. Consideremos una cuaterna de números x, y, z, t; se dice que tal cuaterna está representada por, o simplemente *es*, un punto del espacio tetradimensional R_4. Con más generalidad, un punto del espacio n-dimensional R_n es, por definición, un conjunto ordenado de n números reales $x_1, x_2, ..., x_n$. No importa que no podamos idearnos tal punto. El lenguaje geométrico es igualmente sugestivo para las propiedades algébricas referentes a cuatro o n variables. La razón de esto es que muchas de las propiedades algébricas de las ecuaciones lineales, etc., son esencialmente independientes del número de variables que aparecen o, como podemos decir, de la dimensión del espacio de las variables; p. ej., llamamos «hiperplano» a la totalidad de los puntos $x_1, x_2, ..., x_n$ de un espacio n-dimensional R_n y que satisfacen a la ecuación lineal:

$$L(x_1, x_2, \cdots, x_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b = 0.$$

Por tanto, el problema algebraico fundamental de resolver un sistema de n ecuaciones lineales con n incógnitas:

$$\begin{aligned} L_1(x_1, x_2, \cdots, x_n) &= 0 \\ L_2(x_1, x_2, \cdots, x_n) &= 0 \\ \cdots\cdots\cdots\cdots & \\ L_n(x_1, x_2, \cdots, x_n) &= 0, \end{aligned}$$

se plantea en lenguaje geométrico como el de determinar el punto de intersección de los n hiperplanos $L_1 = 0$, $L_2 = 0$, ..., $L_n = 0$.

La ventaja de este modo geométrico de expresión radica solamente en que hace hincapié sobre ciertas propiedades algebraicas que son independientes de n, *y que pueden imaginarse intuitivamente para* $n \leqslant 3$. En muchas aplicaciones, el uso de tal terminología tiene la ventaja de abreviar, facilitar y dirigir las consideraciones intrínsecamente analí-

ticas. La teoría de la relatividad puede mencionarse como ejemplo donde se ha logrado un importante progreso unificando las coordenadas espaciales x, y, z y la coordenada de tiempo t, en un «suceso» perteneciente a una variedad tetradimensional «espacio-tiempo» de cuaternas de números x, y, z, t. Por la introducción de una geometría no euclídea hiperbólica en este esquema analítico se hace posible describir con gran sencillez muchas situaciones de otro modo complejas. Ventajas análogas se han obtenido en mecánica y en física estadística, así como en campos puramente matemáticos.

He aquí algunos ejemplos tomados de las matemáticas. La totalidad de los círculos del plano constituye una variedad tridimensional, porque un círculo de centro x, y, y radio t puede ser representado por un punto de coordenadas x, y, t. Puesto que el radio de un círculo es un número positivo, la totalidad de los puntos representativos de círculos llena un semiespacio. De la misma manera, el conjunto de todas las esferas del espacio tridimensional ordinario constituye una variedad tetradimensional, ya que cada esfera de centro x, y, z y radio t puede representarse por un punto de coordenadas x, y, z, t. Un cubo en el espacio tridimensional, de arista 2, caras paralelas a los planos coordenados y centro en el origen, está formado por la totalidad de los puntos x_1, x_2, x_3 para los cuales $|x_1| \leqslant 1$, $|x_2| \leqslant 1$, $|x_3| \leqslant 1$. De la misma manera, un «cubo» de arista 2 en un espacio n-dimensional R_n, caras paralelas a los planos coordenados y centro en el origen, se define como la totalidad de los puntos $x_1, x_2, \ldots, x_n$ para los cuales se verifica simultáneamente

$$|x_1| \leqslant 1, \qquad |x_2| \leqslant 1, \cdots, |x_n| \leqslant 1.$$

La «superficie» de este cubo consta de todos los puntos para los que se verifica uno al menos de los signos de igualdad. Los elementos de superficie de dimensión $n - 2$ constan de los puntos para los cuales tienen lugar *dos* al menos de los signos de igualdad, etc.

Ejercicio: Descríbase la superficie de uno de tales cubos en el caso de tres, cuatro y n dimensiones.

***3. Método geométrico o combinatorio.**—Mientras la construcción analítica de la geometría n-dimensional es sencilla y se adapta bien a muchas aplicaciones, hay otro método de proceder que es de carácter puramente geométrico. Éste se basa en una reducción de los datos de n a $n - 1$ dimensiones, lo que nos permitirá definir la geometría de varias dimensiones por un proceso de inducción matemática.

Comencemos con el contorno de un triángulo *ABC* en dos dimensiones. Si se corta la poligonal cerrada por el punto *C* y se hacen girar *AC* y *BC* hasta que vengan sobre la recta *AB*, obtenemos el segmento de la figura 116, en el cual el punto *C* aparece dos veces. Esta figura unidimensional da una representación completa del contorno del triángulo de dos dimensiones. Plegando los dos segmentos *AC* y *BC* en el plano, podemos hacer que los dos puntos *C* coincidan de nuevo. Pero, y esto es el aspecto importante, no precisamos hacer tal plegado. Necesitamos solamente «identificar», o sea, no distinguir entre los dos puntos *C* de la figura 116, incluso aunque no coincidan efectivamente, como entes geométricos en sentido corriente. Podemos dar

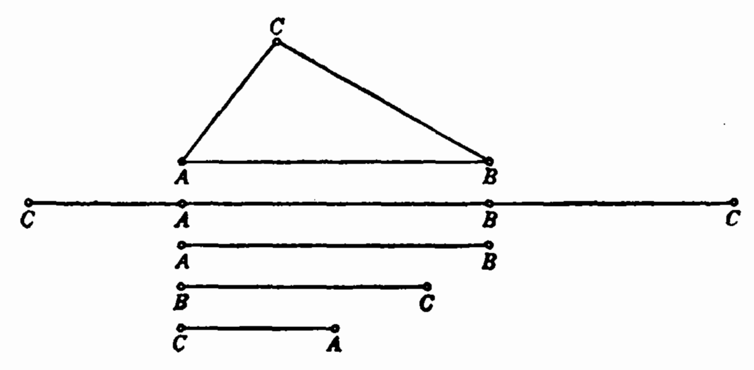


FIG. 116.—Triángulo definido por segmentos con los extremos coordinados.

aún un paso más separando los tres segmentos por los puntos *A* y *B*, para obtener un conjunto de tres segmentos *CA*, *AB* y *BC*, que pueden ser reunidos de nuevo para formar un triángulo «real» sin más que hacer coincidir los pares de puntos identificados. Esta idea de identificar puntos diferentes en un conjunto de segmentos para formar un polígono (en este caso, un triángulo) resulta a veces muy práctica. Si deseamos embarcar una complicada estructura de acero, como la de un puente, la reducimos a simples barras y marcamos con el mismo símbolo aquellos extremos que deben ser unidos cuando la estructura vaya a ser armada en el espacio. El sistema de barras con extremos marcados es completamente equivalente a la estructura espacial. Esta observación sugiere el camino para reducir la superficie bidimensional de un poliedro en un espacio de tres dimensiones a figuras de menor número de dimensiones. Tomemos, p. ej., la superficie de un cubo (Fig. 117); puede ser inmediatamente reducida a un sistema de seis cuadrados planos cuyos segmentos de contorno se identifiquen apro-

piadamente, y en otro paso ulterior se reduce a un sistema de doce segmentos rectilíneos con sus extremos convenientemente identificados.

En general, cualquier poliedro del espacio tridimensional R_3 puede reducirse de esta forma, bien a un sistema de polígonos planos, o bien a un sistema de segmentos rectilíneos.

Ejercicio: Efectúese esta reducción para todos los poliedros regulares (véase página 249).

Queda ahora completamente claro que podemos invertir nuestro razonamiento, *definiendo* un polígono en el plano por un sistema de segmentos rectilíneos, y un poliedro en R_3 por un sistema de polígonos en R_2, o bien, con una mayor reducción, por un sistema de segmentos rectilíneos. Por consiguiente, es natural definir un «poliedro» en el espacio de cuatro dimensiones R_4, mediante un sistema de po-

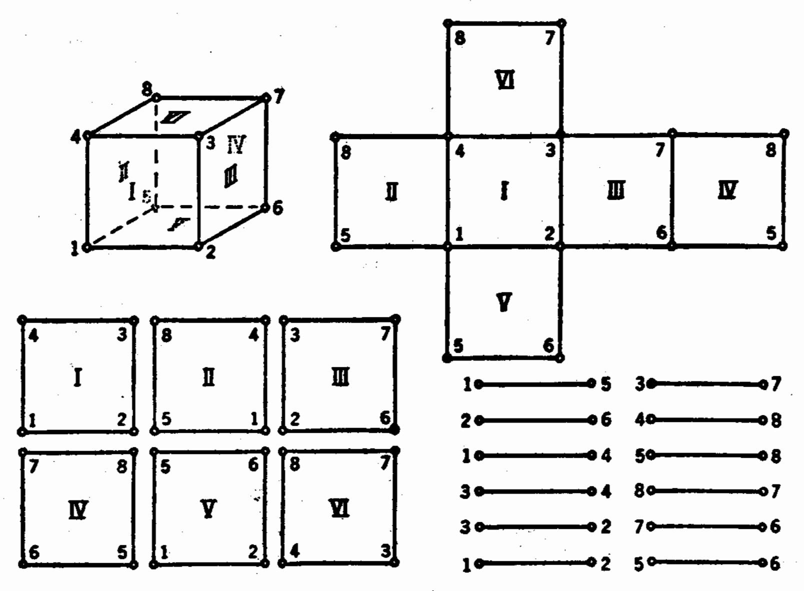


FIG. 117.—Cubo definido por coordinación de vértices y aristas.

liedros en R_3 con apropiada identificación de sus caras bidimensionales; los poliedros de R_5 por sistemas de poliedros en R_4, y así sucesivamente. En definitiva, podemos reducir cualquier poliedro en R_n a un sistema de segmentos rectilíneos.

No es posible desarrollar aquí por completo esta teoría; sólo pueden agregarse algunas observaciones sin demostración. Un cubo en R_4 está limitado por 8 cubos de tres dimensiones, cada uno identificado

con un «vecino» a lo largo de una cara bidimensional. El cubo en R_4 tiene 16 vértices, en cada uno de los cuales se encuentran 4 de las 32 aristas. En R_4 hay 6 poliedros regulares; además del «cubo», hay uno limitado por 5 tetraedros regulares, uno limitado por 16 tetraedros, uno limitado por 24 octaedros, uno limitado por 120 dodecaedros y otro limitado por 600 tetraedros. Para $n > 4$ dimensiones se ha demostrado que sólo son posibles tres poliedros regulares: uno con $n + 1$ vértices, limitado por $n + 1$ poliedros en R_{n-1}, con n aristas

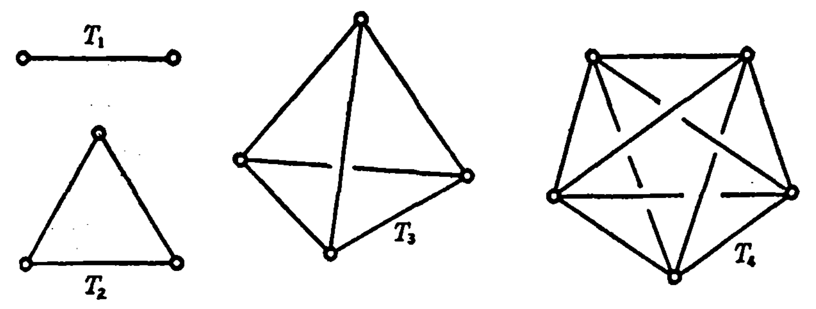


FIG. 118.—Las figuras geométricas más sencillas en 1, 2, 3 y 4 dimensiones.

de $(n - 2)$ dimensiones; uno con 2^n vértices, limitado por $2n$ poliedros en R_{n-1}, con $2n - 2$ aristas, y uno con $2n$ vértices, limitado por 2^n poliedros, de n caras, en R_{n-1}.

Ejercicio: Compárese la definición del cubo en R_4, dada en la página 242, con la definición que acabamos de dar ahora, y pruébese que la definición «analítica» de la superficie del cubo citado es equivalente a la definición «combinatoria» que se acaba de dar.

Desde el punto de vista estructural, o «combinatorio», las figuras geométricas más sencillas de dimensiones 0, 1, 2, 3, son el punto, el segmento, el triángulo y el tetraedro, respectivamente. Con objeto de dar una notación uniforme designemos estas figuras con los símbolos T_0, T_1, T_2, T_3, respectivamente. (Los subíndices señalan la dimensión.)

La estructura de cada una de estas figuras se describe diciendo que cada T_n contiene $n + 1$ vértices y que cada subconjunto de $i + 1$ vértices de un $T_n (i = 0, 1, ..., n)$ determina un T_i; p. ej., el tetraedro tridimensional T_3 contiene 4 vértices, 6 segmentos y 4 triángulos.

Es evidente la forma de proceder: Definimos un «tetraedro» tetradimensional T_4 como un conjunto de 5 vértices tales que cada subconjunto de 4 vértices determina un T_3, cada subconjunto de 3 vértices determina un T_2, etc. El diagrama esquemático de T_4 se muestra

en la figura 118. Vemos que T_4 contiene 5 vértices, 10 segmentos, 10 triángulos y 5 tetraedros.

La generalización para n dimensiones es inmediata. Por la teoría de las combinaciones se sabe que hay exactamente $C_i^r = \frac{r!}{i!(r-i)!}$ subconjuntos diferentes, de i objetos cada uno, que pueden formarse con un conjunto dado de r objetos. Luego un «tetraedro» n-dimensional contiene:

$$C_1^{n+1} = n + 1 \qquad \text{vértices} \qquad (T_0),$$

$$C_2^{n+1} = \frac{(n+1)!}{2!(n-1)!} \qquad \text{segmentos} \qquad (T_1),$$

$$C_3^{n+1} = \frac{(n+1)!}{3!(n-2)!} \qquad \text{triángulos} \qquad (T_2),$$

$$C_4^{n+1} = \frac{(n+1)!}{4!(n-3)!} \qquad T_3,$$

. . . .

$$C_{n+1}^{n+1} = 1 \qquad T_n.$$

Ejercicio: Dibújese un diagrama de T_5 y determínese el número de los T_i diferentes que contiene, para $i = 0, 1, 2, 3, 4, 5$.

CAPÍTULO V

TOPOLOGÍA

Introducción.—A mediados del siglo XIX comenzó un desarrollo enteramente nuevo de la geometría, que pronto se convirtió en una de las fuerzas más potentes de la matemática moderna. La nueva disciplina, llamada *análisis situs* o *topología*, estudia las propiedades de las figuras geométricas que subsisten aun si esas figuras se someten a deformaciones tan radicales que las hagan perder todas sus propiedades métricas y proyectivas.

Uno de los grandes geómetras de esa época fue A. F. Moebius (1790-1868), un hombre cuya falta de seguridad en sí mismo le llevó al puesto de insignificante astrónomo en un observatorio de importancia secundaria de Alemania. A la edad de sesenta y ocho años sometió a la Academia de París una memoria sobre superficies de «una sola cara», que contenía algunos de los hechos más sorprendentes de este nuevo tipo de geometría. Al igual que otras importantes contribuciones anteriores, su trabajo permaneció sepultado varios años en los archivos de la Academia, hasta que por fin lo publicó su autor. Independientemente de Moebius, el astrónomo J. B. Listing (1808-1882), de Gotinga, hizo descubrimientos análogos, y a sugerencia de Gauss publicó en 1847 un pequeño libro, *Vorstudien zur Topologie.* Cuando Bernhard Riemann (1826-1866) llegó a Gotinga como estudiante, encontró la atmósfera matemática de esa ciudad universitaria llena de ansioso interés por estas nuevas y extrañas ideas geométricas. Pronto se dió cuenta de que allí estaba la clave para comprender las propiedades más profundas de las funciones analíticas de una variable compleja. Nada, quizá, ha dado más ímpetu al posterior desarrollo de la topología que la formidable estructura de la teoría de funciones de Riemann, en la cual los conceptos topológicos son absolutamente fundamentales.

Al principio, la novedad de los métodos en el reciente campo no dejó tiempo a los matemáticos para presentar sus resultados en la tradicional forma axiomática de la geometría elemental. En lugar de ello, los primeros investigadores, como Poincaré, se vieron forzados a confiar ampliamente en la intuición geométrica. Aún hoy, un estudiante de topología encontrará que si exagera la insistencia en el rigor de la presentación puede perder de vista el contenido geomé-

trico esencial entre la masa de detalles formales. Sin embargo, es un gran mérito de los trabajos recientes el haber incluído la topología dentro del marco de la matemática rigurosa, donde la intuición sigue siendo la fuente, pero no la última razón de validez de la verdad. Durante este proceso, comenzado por L. E. J. Brouwer, la importancia de la topología para casi toda la matemática se ha ido incrementando. Matemáticos americanos, en particular O. Veblen, J. W. Alexander y S. Lefschetz, han aportado importantes contribuciones al tema.

Aunque la topología es, en definitiva, una creación de los últimos cien años, hubo unos pocos descubrimientos aislados anteriores, que después encontraron su lugar en el moderno desarrollo sistemático. Sin duda, el más importante de ellos es una fórmula que relaciona el número de vértices, aristas y caras de un poliedro simple, observada ya en 1640 por Descartes, y redescubierta y utilizada por Euler en 1752. El típico carácter de esta relación como teorema topológico se hizo evidente mucho más tarde, después de que Poincaré reconoció «la fórmula de Euler» y sus generalizaciones como uno de los teoremas centrales de la topología. Así, por razones tanto históricas como intrínsecas, iniciaremos nuestro estudio de la topología con la fórmula de Euler. Puesto que el ideal del rigor perfecto no es ni necesario ni deseable durante los primeros pasos en un campo no familiar, no dudaremos en apelar de cuando en cuando a la intuición geométrica del lector.

I. FÓRMULA DE EULER PARA LOS POLIEDROS

Aunque el estudio de los poliedros ocupó un lugar privilegiado en la geometría griega, cupo a Descartes y a Euler el descubrimiento del siguiente hecho: en un poliedro simple, si se designa por V el número de vértices, por A el de aristas y por C el número de caras, se verifica

$$V - A + C = 2. \qquad [1]$$

Por *poliedro* se entiende un sólido cuya superficie consta de un cierto número de caras poligonales. En el caso de los poliedros regulares todos los polígonos son congruentes y todos los ángulos en los vértices son iguales. Un poliedro es *simple* si no hay en él «agujeros»; o sea, si su superficie puede ser deformada con continuidad hasta transformarse en la superficie de una esfera. La figura 120 muestra un poliedro simple no regular, mientras la figura 121 representa un poliedro que no es simple.

El lector puede comprobar el hecho de que la fórmula de Euler

se cumple para los poliedros simples de las figuras 119 y 120, pero no para el poliedro de la figura 121.

Para demostrar la fórmula de Euler, imaginemos que el poliedro simple dado es hueco y su superficie de caucho. Si separamos una de las caras, podremos deformar la superficie restante hasta extenderla sobre un plano. Por supuesto que las áreas de las caras y sus án-

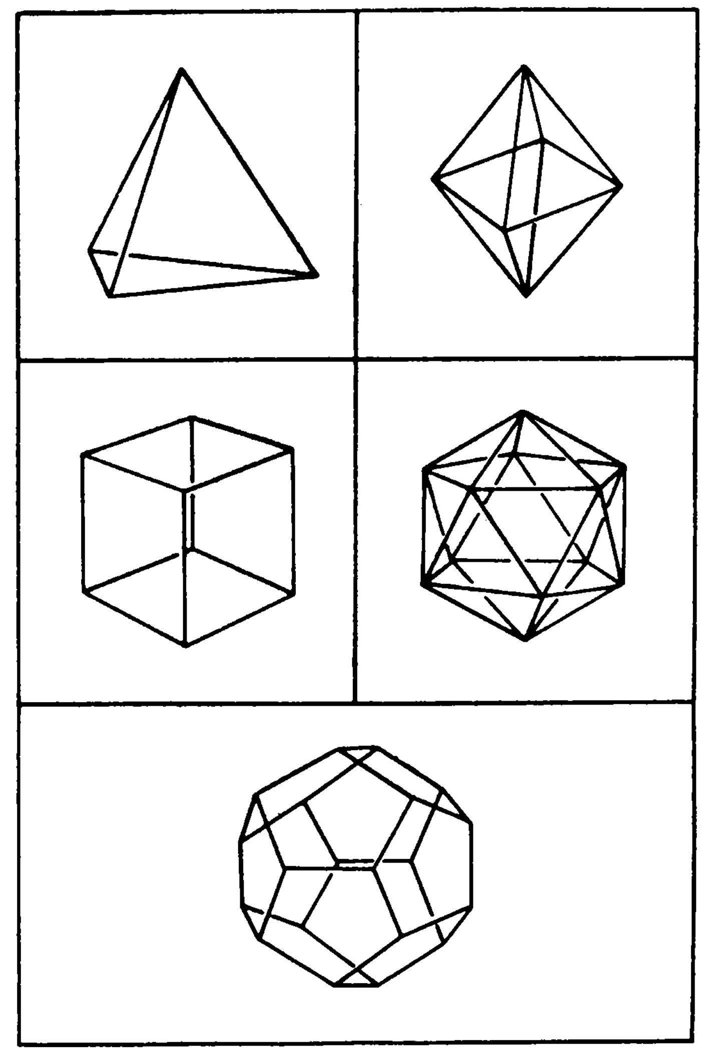


FIG. 119.—Los poliedros regulares.

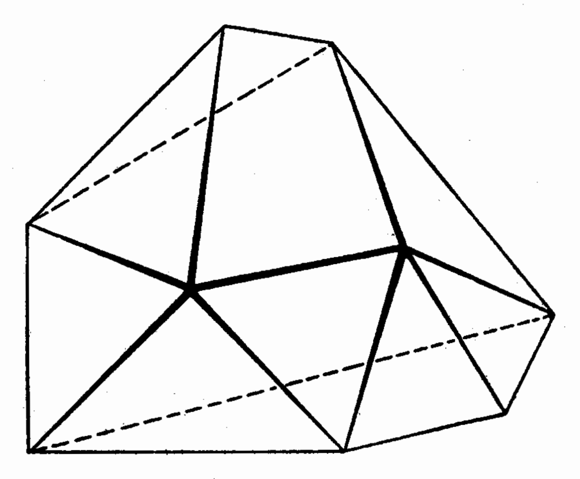


FIG. 120.—Poliedro simple. $V - A + C = 9 - 18 + 11 = 2$.

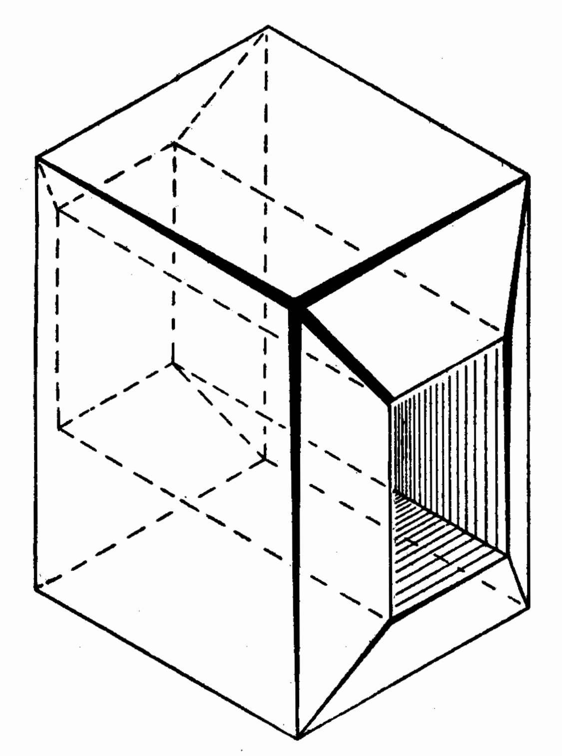


FIG 121.—Poliedro no simple. $V - A + C = 16 - 32 + 16 = 0$.

gulos se alterarán en este proceso, pero la red plana de vértices y aristas contendrá el mismo número de unos y otros que el poliedro original, en tanto que el número de polígonos es inferior en uno al del primitivo poliedro del cual hemos suprimido una cara. Vamos a probar que para la red plana obtenida $V - A + C = 1$, de forma que si se tiene en cuenta la cara suprimida resulta $V - A + C = 2$, para el poliedro dado.

Comencemos por *triangular* la red plana del siguiente modo: En cualquier polígono de la red que no sea triángulo trazamos una diagonal, la cual incrementa ambos números A y C en 1, de forma que

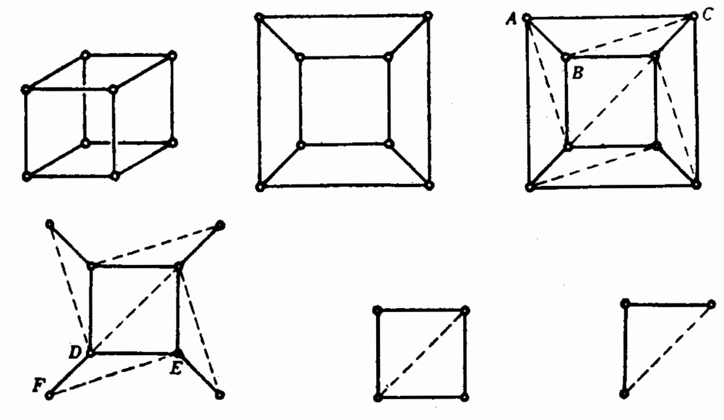


FIG. 122.—Demostración de la fórmula de Euler.

el valor de $V - A + C$ se conserva. Continuamos el trazado de diagonales que unen pares de puntos (Fig. 122) hasta que la figura consta sólo de triángulos, lo que puede ocurrir ya desde un principio. En la red triangulada, $V - A + C$ conserva el mismo valor que tenía antes de la triangulación, puesto que el trazado de diagonales no lo ha alterado. Algunos de los triángulos tienen sus lados en la frontera de la red plana, y de éstos, algunos—tal como el ABC—tienen sólo un lado formando parte de dicha frontera, mientras que otros pueden tener dos. Elegimos uno de estos triángulos frontera y suprimimos los lados que no sean comunes con algún otro triángulo; así, del ABC quitamos el lado AC y descontamos la cara correspondiente, pero conservamos los vértices A, B, C y los otros dos lados AB y BC, en tanto que del triángulo DEF suprimimos la cara, los dos lados DF y FE y el vértice F. La supresión de un triángulo del tipo ABC disminuye los números A y C en 1, mientras V queda invariable, de

forma que $V - A + C$ permanece inalterada. La supresión de un triángulo del tipo DEF hace disminuir V en 1, A en 2 y C en 1, de manera que $V - A + C$ continúa invariable. Mediante una adecuada elección de la sucesión de estas operaciones podemos ir separando triángulos que tengan algún lado en la frontera (la cual varía, por supuesto, con cada supresión), hasta dejar finalmente un solo triángulo, con tres vértices, tres lados y una cara. Para esta red simplificada $V - A + C = 3 - 3 + 1 = 1$, y como hemos visto que el proceso no alteraba el valor de $V - A + C$, también en la red plana inicial será $V - A + C = 1$, y lo mismo para el poliedro del que habíamos suprimido una cara. En conclusión, en el poliedro inicial completo $V - A + C = 2$, con lo cual queda demostrada la fórmula de Euler (véanse los problemas 56 y 57 del Apéndice).

Basándonos en la fórmula de Euler es fácil demostrar que no existen más que cinco poliedros regulares. Supongamos, en efecto, que un poliedro regular tiene C caras, siendo cada una un polígono regular de n lados, y que en cada vértice concurren r aristas. Si contamos las aristas por el número de caras o el de vértices, se tiene

$$nC = 2A, \qquad [2]$$

pues cada arista es común a dos caras y está por ello contada dos veces en el producto nC; además,

$$rV = 2A, \qquad [3]$$

ya que cada arista tiene dos vértices. De [1], obtenemos la ecuación

$$\frac{2A}{n} + \frac{2A}{r} - A = 2,$$

o bien

$$\frac{1}{n} + \frac{1}{r} = \frac{1}{2} + \frac{1}{A} \qquad [4]$$

Debemos comenzar con $n > 3$ y $r > 3$, ya que un polígono tiene al menos tres lados, y por lo menos tres aristas concurren en cada vértice del poliedro. Pero n y r no pueden ser *mayores* que 3, pues el primer miembro de la ecuación [4] no excedería a $^1/_2$, lo que es imposible, ya que A es entero positivo. Por tanto, veamos qué valores puede tener r para $n = 3$ y cuáles son los de n para $r = 3$. La totalidad de los poliedros obtenidos en estos dos casos nos da el número de poliedros regulares posibles.

Para $n = 3$, la ecuación [4] se transforma en

$$\frac{1}{r} - \frac{1}{6} = \frac{1}{A},$$

de donde r puede ser igual a 3, 4 ó 5 (6 ó un número mayor queda excluído, pues $1/A$ es siempre positivo). Para estos valores de n y r hallamos $A = 6$, 12 ó 30,

que corresponden, respectivamente, al tetraedro, octaedro e icosaedro. Análogamente, para $r = 3$ se obtiene la ecuación

$$\frac{1}{n} - \frac{1}{6} = \frac{1}{A},$$

de la cual resulta que $n = 3$, 4 ó 5, y $A = 6$, 12 ó 30, respectivamente. Estos valores corresponden al tetraedro, cubo y dodecaedro. Si sustituímos estos valores de n, r y A en las ecuaciones [2] y [3], obtenemos los números de vértices y caras de los respectivos poliedros.

II. PROPIEDADES TOPOLÓGICAS DE LAS FIGURAS

1. Propiedades topológicas.—Hemos visto que la fórmula de Euler se verifica para todo poliedro simple, pero la validez de esta fórmula no queda restringida, ni mucho menos, a los poliedros de la geometría elemental, de caras planas y aristas rectas; la demostración que acabamos de dar se aplicaría igualmente a poliedros simples de caras y aristas curvas, o a cualquier subdivisión de la superficie de una esfera en regiones limitadas por arcos de curvas. Por otra parte, si imaginamos la superficie del poliedro o de la esfera constituída por una delgada lámina de caucho, la fórmula de Euler todavía conserva su validez si la superficie se deforma doblando y estirando el caucho hasta darle otra cualquier forma, en tanto este proceso de deformación no produzca desgarramiento. Esto se debe a que la fórmula se refiere sólo a los *números* de vértices, aristas y caras y no a longitudes, áreas, razones dobles o cualquier otro concepto ordinario de la geometría elemental o proyectiva.

Recordemos que la geometría elemental tiene que ver con las magnitudes (longitud, ángulo y área) que quedan invariables en los movimientos rígidos, en tanto que la geometría proyectiva trata con conceptos (punto, recta, incidencia y razón doble) que quedan invariables en el grupo más amplio de las transformaciones proyectivas. Pero los movimientos rígidos y las proyecciones son casos muy especiales de las llamadas *transformaciones topológicas*. Una transformación topológica de una figura geométrica A en otra A' está dada por cualquier correspondencia

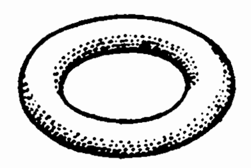
$$p \longleftrightarrow p'$$

entre los puntos p de A y los p' de A', caracterizada por las dos propiedades siguientes:

1) *La correspondencia es biunívoca.* Esto significa que a cada punto p de A corresponde un solo punto p' de A', y recíprocamente.

2) *La correspondencia es continua en ambos sentidos.* Esto quiere decir que si tomamos dos puntos cualesquiera p, q de A y movemos p de forma que su distancia al punto q tienda a cero, la distancia entre los puntos correspondientes p'. q' de A' también tiende a cero, y recíprocamente.

Toda propiedad de una figura geométrica A que también se cumpla para cualquier figura en que A pueda transformarse por una transformación topológica, se llama una *propiedad topológica* de A, y



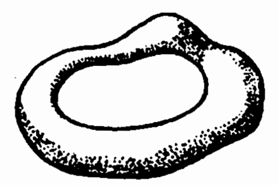
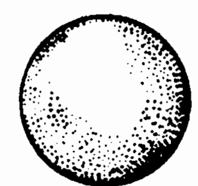


FIG. 123.—Superficies topológicamente equivalentes.



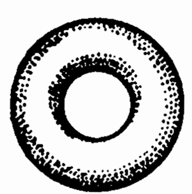


FIG. 124.—Superficies no equivalentes topológicamente.

topología es la rama de la geometría que se ocupa sólo de las propiedades topológicas de las figuras. Imaginemos una figura copiada «libremente» por un consciente pero inexperto dibujante, que deforme las rectas en curvas y altere los ángulos, distancias y áreas; pues bien, aunque las propiedades métricas y proyectivas de la figura original se hayan perdido, sus propiedades topológicas permanecerán idénticas.

Los ejemplos más intuitivos de transformaciones topológicas generales son las *deformaciones*. Imaginemos una figura, como una superficie esférica o un triángulo, fabricada con una delgada lámina de caucho o dibujada sobre ella, la cual es luego estirada o retorcida de cualquier forma, sin rasgarla y sin hacer coincidir puntos distintos. (Hacer coincidir dos puntos distintos violaría la condición 1. Rasgar la lámina de caucho violaría la condición 2, puesto que dos puntos de la figura original que tienden a coincidir provenientes de lados opuestos de la línea a lo largo de la cual se ha rasgado la lámina, no tenderán a coincidir en la figura rasgada.) La posición final de la figura

debe ser, pues, una imagen topológica de la original. Un triángulo puede ser deformado hasta obtener otro triángulo, un círculo (figuras 123 y 124) o una elipse, y, por tanto, estas figuras tienen exactamente las mismas propiedades topológicas; pero no se puede deformar un círculo de manera que resulte un segmento rectilíneo, ni la superficie de una esfera puede dar lugar a la superficie interior de un tubo.

El concepto general de transformación topológica es más amplio que el concepto de deformación. Por ejemplo, si una figura se rasga durante una deformación y seguidamente se unen los bordes del corte después de deformada, de igual manera que antes, el proceso define aún una transformación topológica de la figura original, aunque ya no es una deformación. Así las dos curvas de la figura 134 son topológicamente equivalentes entre sí y lo son también a una circunferencia, puesto que pueden cortarse, enderezarse, y luego unir ambos extremos. Pero es imposible deformar una curva en la otra o en una circunferencia, sin antes efectuar el corte de la curva.

Las propiedades topológicas de las figuras (tales como la dada por el teorema de Euler y otras que serán discutidas en esta sección)

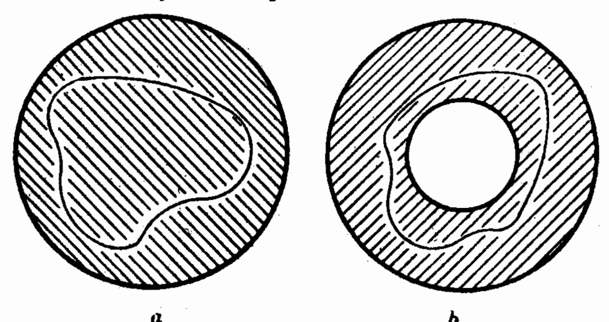


FIG. 125.—Conexión simple y doble.

tienen gran interés e importancia en muchas investigaciones matemáticas. Son, en cierto sentido, las más profundas y fundamentales de todas las propiedades geométricas, puesto que persisten después de realizados los cambios de forma más radicales.

2. Conexión.—Como nuevo ejemplo de dos figuras que no son topológicamente equivalentes, podemos considerar los dominios planos de la figura 125. El primero de ellos consta de todos los puntos interiores a un círculo, mientras que el segundo lo constituyen todos los puntos situados entre dos circunferencias concéntricas. Cualquier curva cerrada contenida en el dominio *a* puede deformarse o «con-

traerse» continuamente alrededor de un único punto *dentro del dominio*. Un dominio con esta propiedad se dice *simplemente conexo*. El dominio *b* no es simplemente conexo; p. ej., una circunferencia concéntrica con las dos circunferencias que sirven de frontera no puede deformarse hasta reducirse a un único punto interior al dominio, puesto que durante este proceso la curva debería pasar necesariamente por el centro de las circunferencias, que no es un punto del dominio. Un dominio que no es simplemente conexo se llama *múltiplemente conexo*. Si el dominio múltiplemente conexo *b* se corta a lo largo de un radio, como en la figura 126, el dominio resultante es simplemente conexo.

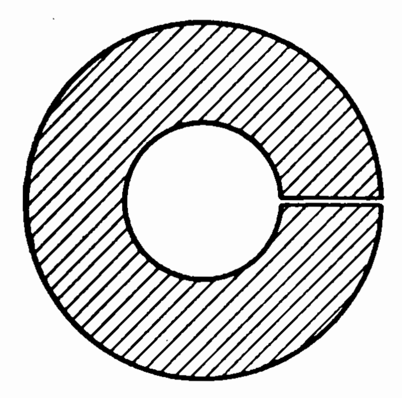


FIG. 126.—Corte de un dominio doblemente conexo para lograr uno simplemente conexo.

Con mayor generalidad, podemos construir dominios con dos, tres,

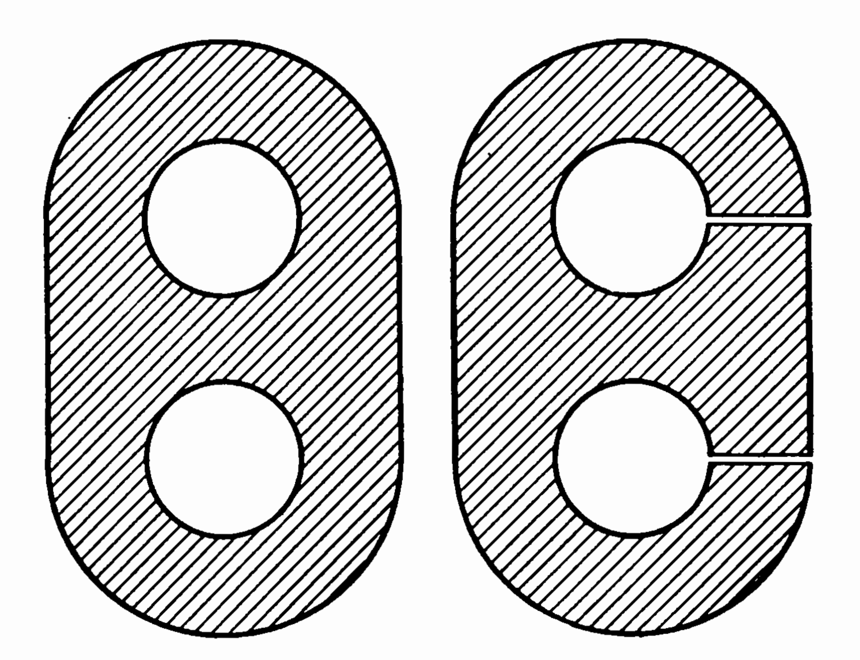


FIG. 127.—Reducción de un dominio triplemente conexo.

o más «agujeros», como el de la figura 127. Para convertir este dominio en otro simplemente conexo, son necesarios dos cortes. Si para convertir un dominio dado D múltiplemente conexo en otro simplemente conexo se necesitan $n - 1$ cortes que no se atraviesen y que unan un borde con otro, entonces el dominio D se dice que tiene orden de conexión n. El orden de conexión de un dominio en el plano es un importante invariante topológico del dominio.

III. OTROS EJEMPLOS DE TEOREMAS TOPOLÓGICOS

1. **El teorema de la curva de Jordan.**—Una curva cerrada y simple (o sea, que no se corta a sí misma) está dibujada en el plano. ¿Qué propiedad de esta figura subsiste cuando el plano se considera como una lámina de caucho que puede deformarse de cualquier modo? La longitud de la curva y el área que encierra pueden alterarse con una deformación; pero hay una propiedad topológica de la configuración que es tan simple que puede parecer trivial: *una curva simple y cerrada* C *de un plano divide a éste en dos dominios, uno interior y otro exterior.* Esto significa que los puntos del plano se dividen en dos clases—A, el exterior de la curva, y B, el interior—, tales que cualquier par de puntos de la misma clase pueden unirse por medio de una curva que no corta a C, mientras que cualquier curva que una dos puntos pertenecientes a clases distintas debe cortar a C. Esta aserción es evidentemente cierta para una circunferencia o una elipse; pero esta evidencia disminuye notablemente si se considera una curva complicada tal como el intrincado polígono de la figura 128.

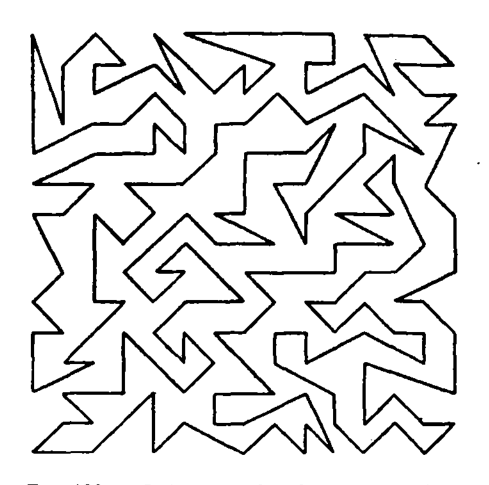


FIG. 128.—¿Qué puntos del plano son interiores a este polígono?

Este teorema fué establecido por primera vez por Camille Jordan (1838-1922), en su famoso *Cours d'Analyse*, en el cual toda una generación de matemáticos aprendió el concepto moderno de rigor en aná-

lisis. Aunque parezca extraño, la demostración dada por Jordan no era ni corta ni sencilla, y la sorpresa resulta aún mayor si se advierte que dicha demostración era falsa y que se necesitó un considerable esfuerzo para llenar las lagunas de su razonamiento. La primera demostración rigurosa del teorema resultó muy complicada y difícil de entender, aun para muchos matemáticos bien entrenados. Sólo recientemente se han dado demostraciones relativamente sencillas. Una razón de la dificultad estriba en la generalidad del concepto de «curva simple cerrada», que no sólo incluye al conjunto de polígonos y curvas «uniformes», sino también a todas aquellas curvas, imágenes topológicas de una circunferencia. Por otra parte, muchos conceptos como «interior», «exterior», etc., que son tan claros para la intuición, deben precisarse antes de poder dar una demostración rigurosa. Es de gran importancia teórica analizar tales conceptos en su más amplia generalidad, y gran parte de la topología moderna se consagra a este fin. Pero no debe olvidarse nunca que en la gran mayoría de los casos que se presentan en el estudio de los fenómenos geométricos concretos resulta innecesario manejar conceptos cuya excesiva generalidad cree dificultades adicionales. Así, p. ej., el teorema de la curva de Jordan puede demostrarse con relativa facilidad para curvas de «comportamiento» razonable, como polígonos o curvas con tangente que varía con continuidad, que son las que aparecen en la mayoría de los problemas importantes. Demostraremos el teorema para los polígonos en el apéndice de este capítulo.

2. El problema de los cuatro colores.—A partir del ejemplo del teorema de la curva de Jordan, podría suponerse que la topología se ocupa de dar demostraciones rigurosas de ciertas aserciones obvias, de las cuales ninguna persona en su sano juicio dudaría. Por el contrario, hay muchas cuestiones topológicas, algunas de ellas de forma muy simple, para las cuales la intuición no da respuesta satisfactoria. Un ejemplo de este tipo es el famoso «problema de los cuatro colores».

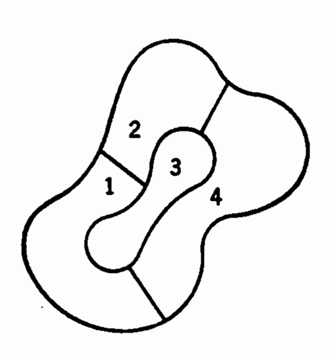


FIG. 129.—Iluminación de un mapa.

Al iluminar un mapa geográfico, se acostumbra asignar diferentes colores a dos países cualesquiera que tienen una porción de frontera común. Se ha encontrado empíricamente que cualquier mapa, indepen-

dientemente del número de países que contenga y de cómo estén éstos situados, puede iluminarse utilizando sólo cuatro colores diferentes. Es fácil ver que un número menor de colores no es suficiente para todos los casos. La figura 129 nos muestra una isla que, en efecto, no puede iluminarse con menos de cuatro colores, ya que pertenece a cuatro naciones, cada una de las cuales tiene frontera con las otras tres.

El hecho de que no se haya encontrado ningún mapa que requiera más de cuatro colores sugiere el siguiente teorema matemático: *Para cualquier subdivisión del plano en regiones que no se solapen, es siempre posible señalar las regiones con uno de los números* 1, 2, 3, 4, *de tal modo que a dos regiones adyacentes no se les asigne nunca el mismo número.* Por regiones «adyacentes» entendemos regiones con todo un arco de frontera común: dos regiones que sólo tengan un punto o un número finito de puntos comunes (como los estados de Colorado y Arizona) no se llamarán adyacentes, puesto que no puede surgir ninguna confusión si se iluminan con el mismo color.

El problema de demostrar este teorema parece haber sido propuesto primeramente por Moebius en 1840, después por DeMorgan en 1850, y de nuevo por Cayley en 1878. Kempe publicó una «demostración» en 1879; pero, en 1890, Heawood encontró un error en el razonamiento de Kempe. Mediante la revisión de la demostración de Kempe, Heawood pudo demostrar que *cinco* colores son siempre suficientes. (Una demostración del teorema de los cinco colores se da en el apéndice de este capítulo.) A pesar de los esfuerzos de muchos matemáticos insignes, la cuestión sigue aún esencialmente estancada en este modesto resultado: se ha *probado* que cinco colores bastan para todos los mapas y se *supone* que cuatro colores son también suficientes. Pero, como en el caso del famoso teorema de Fermat (véase pág. 50), no se ha logrado ni una demostración de esta conjetura ni un contraejemplo, por lo cual sigue siendo éste uno de los grandes problemas no resueltos de la matemática. El teorema de los cuatro colores fué realmente demostrado para todos los mapas que contienen menos de 38 regiones. En vista de este hecho, parece que incluso si el teorema general es falso, no podrá demostrarse esta falsedad mediante un ejemplo sencillo.

En el problema de los cuatro colores los mapas pueden dibujarse, bien en el plano o sobre la superficie de una esfera. Los dos casos son equivalentes; todo mapa sobre la esfera puede representarse sobre el plano haciendo un pequeño agujero interior a una de las regiones A y deformando la superficie resultante hasta hacerla plana, como en

la demostración del teorema de Euler. El mapa resultante en el plano, puede ser el de una «isla» constituída por las regiones restantes, rodeada por «mar», que es la región A. Recíprocamente, invirtiendo este proceso, cualquier mapa plano puede representarse sobre una esfera. Podemos, por tanto, limitarnos a los mapas sobre la esfera. Además, dado que las deformaciones de las regiones y de sus fronteras no afectan al problema, podemos suponer que la frontera de cada región es un polígono simple cerrado, formado por arcos circulares. Aun así «regularizado», el problema permanece sin resolver; las dificultades en este caso, por oposición a las inherentes al teorema de la curva de Jordan, no residen en la generalidad de los conceptos de región y curva.

Un hecho notable, relativo al problema de los cuatro colores, es el de que para superficies más complicadas que el plano y la esfera, los teoremas correspondientes han sido efectivamente demostrados, o sea que, por paradójico que parezca, el análisis de superficies geométricas más complicadas aparece a este respecto más fácil que el de los casos más sencillos. Por ejemplo, sobre la superficie de un toro (véase figura 123), cuya forma es la de un buñuelo o de un neumático hinchado, se ha demostrado que cualquier mapa puede iluminarse utilizando siete colores, mientras pueden construirse mapas que contienen siete regiones, cada una de las cuales es fronteriza a las otras seis.

***3. El concepto de dimensión.**—El concepto de dimensión no presenta grandes dificultades mientras se trata de figuras geométricas sencillas, tales como puntos, rectas, triángulos y poliedros. Un punto único o un conjunto *finito* de puntos tiene dimensión cero; un segmento rectilíneo es unidimensional, y la superficie de un triángulo o la de una esfera son bidimensionales. El conjunto de los puntos de un cubo sólido es tridimensional. Pero, en cuanto se trata de extender este concepto a conjuntos más generales de puntos, surge la necesidad de dar una definición precisa. ¿Qué dimensión deberá asignarse al conjunto de puntos R formado por todos los puntos del eje x, de abscisa *racional*? El conjunto de los puntos racionales es denso sobre el segmento, por lo que podría considerarse unidimensional como el propio segmento. Por otra parte, existen algunas irracionales entre cualquier par de puntos racionales, como

Fig. 130.—Conjunto de puntos de Cantor.

sucede entre dos puntos cualesquiera de un conjunto finito de puntos, de forma que cabe considerar también que la dimensión del conjunto R es cero.

Se plantea una cuestión aún más espinosa al tratar de asignar una dimensión al siguiente y curioso conjunto de puntos, que Cantor fué el primero en considerar. Suprímase en el segmento unidad la tercera parte central, que se compone de todos los puntos tales que $1/3 < x < 2/3$. Llamemos C_1 al conjunto de puntos restantes. Suprímase ahora en cada uno de los dos segmentos en que está dividido C_1 el tercio central, llamando C_2 al conjunto restante. Repítase este proceso, suprimiendo el tercio central de cada uno de los cuatro intervalos de C_2, con la que se obtiene un nuevo conjunto C_3, y prosígase de esta manera formando los conjuntos $C_4, C_5, C_6, \cdots$. Llamemos C al conjunto de puntos del segmento unidad que quedan después de haber suprimido todos estos intervalos; es decir, C es el conjunto de todos los puntos comunes a todos los conjuntos de la sucesión indefinida $C_1, C_2, C_3, \cdots$. Puesto que se suprimió un intervalo de longitud 1/3 en el primer paso del proceso, dos intervalos de longitud $1/3^2$ cada uno en el segundo, etcétera, la longitud total de los segmentos suprimidos es

$$1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{3^2} + 2^2 \cdot \frac{1}{3^3} + \cdots = \frac{1}{3}\left[1 + \left(\frac{2}{3}\right) + \left(\frac{2}{3}\right)^2 + \cdots\right]$$

El paréntesis es una serie geométrica, cuya suma es $1/(1 - 2/3) = 3$, de donde se deduce que la longitud total de los segmentos suprimidos es 1. Sin embargo, todavía quedan puntos en el conjunto C; p. ej., 1/3, 2/3, 1/9, 2/9, 7/9, 8/9, $\cdots$, que dividen en tres partes a los segmentos sucesivos. No es difícil ver que C se compone precisamente de todos aquellos puntos x, cuyo desarrollo en fracción indefinida en el sistema de base 3, puede escribirse en la forma

$$x = \frac{a_1}{3} + \frac{a_2}{3^2} + \frac{a_3}{3^3} + \cdots + \frac{a_n}{3^n} + \cdots,$$

donde cada uno de los a_i es 0 ó 2, mientras que el desarrollo análogo de cualquiera de los números suprimidos contendría al menos uno de los a_i igual a 1.

¿Cuál será la dimensión de este conjunto C? El proceso diagonal, utilizado para demostrar la no numerabilidad del conjunto de todos los números reales, puede modificarse de tal manera que proporcione el mismo resultado para el conjunto C. En consecuencia, parecerá que el conjunto C es unidimensional; sin embargo, C no contiene ningún intervalo completo, por pequeño que se suponga, por lo que también podría pensarse que tiene la dimensión cero, como cualquier conjunto finito de puntos. Dentro del mismo orden de ideas, cabe preguntarse si el conjunto de puntos del plano obtenido levantando un segmento perpendicular de longitud unidad sobre todo punto racional o sobre cada punto del conjunto de Cantor C, debe considerarse de dimensión uno o dos.

En 1912, Poincaré llamó la atención acerca de la necesidad de analizar más profundamente y de dar una definición precisa del concepto de dimensión. Poincaré observó que la recta es unidimensional, debido a que podemos separar dos puntos cualesquiera de ella, cortándola en un solo punto (conjunto de dimensión cero), mientras el plano es bidimensional, porque para separar dos cualesquiera de sus puntos debemos cortarlo a lo largo de toda una curva cerrada (conjunto de dimensión 1). Esto nos sugiere la naturaleza inductiva de la dimensión:

un espacio es n-dimensional, si se pueden separar dos puntos cualesquiera de él suprimiendo un subconjunto de $(n - 1)$ dimensiones, mientras no es siempre posible obtener el mismo resultado suprimiendo un conjunto de menos dimensiones. Implícitamente, los *Elementos* de Euclides contienen una definición inductiva del concepto de dimensión, donde se dice que una figura unidimensional es aquella cuya frontera está compuesta de puntos; bidimensional si su frontera está formada por curvas, y tridimensional, aquella figura cuya frontera se compone de superficies.

En los últimos años se ha desarrollado un extensa teoría del concepto de dimensión. Una de las definiciones de dimensión comienza precisando el concepto de «conjunto de puntos de dimensión 0». Cualquier conjunto *finito* de puntos tiene la propiedad de que es posible encerrar cada punto del conjunto en una región del espacio arbitrariamente pequeña, que no contiene ningún otro punto del conjunto en su frontera. Esta propiedad es la adoptada actualmente como definición del conjunto de dimensión cero y se conviene en decir que un conjunto vacío (que no contiene ningún punto) tiene dimensión -1. Entonces, un conjunto de puntos S tiene dimensión 0, si no es de dimensión -1 (es decir, si S contiene al menos un punto), y si cada punto de S puede encerrarse en una región arbitrariamente pequeña cuya frontera corte a S en un conjunto de dimensión -1 (es decir, no contenga ningún punto de S); p. ej., el conjunto de los puntos racionales de la recta tiene dimensión cero, puesto que todo punto racional puede ser centro de un intervalo arbitrariamente pequeño, cuyos extremos sean irracionales. Se ve también que el conjunto de Cantor, C, tiene dimensión cero, ya que, análogamente al conjunto de los puntos racionales, se obtiene eliminando un conjunto denso de puntos de la recta.

Hasta ahora hemos definido solamente los conceptos de dimensión 0 y -1; la definición de dimensión 1 resulta inmediata: un conjunto S de puntos es de dimensión 1, si no es de dimensión cero o -1, y si además cada punto de S puede encerrarse en una región arbitrariamente pequeña y tal que su frontera corte a S en un conjunto de dimensión cero. Un segmento rectilíneo tiene esta propiedad, puesto que la frontera de cualquier intervalo es un par de puntos, que es un conjunto de dimensión cero, de acuerdo con la definición anterior. Por otra parte, si se procede análogamente, podemos definir sucesivamente los conjuntos de dimensión 2, 3, 4, 5, ... , cada uno de los cuales se apoya en la definición precedente. Así, un conjunto S tendrá dimensión n si no es de dimensión menor, y si cada punto de S puede encerrarse en una región arbitrariamente pequeña, tal que su frontera corte a S en un conjunto de dimensión $n-1$; p. ej., el plano es de dimensión 2, puesto que cada punto del mismo puede encerrarse dentro de un círculo arbitrariamente pequeño, cuya circunferencia es de dimensión 1[1]. Ningún conjunto de puntos en el espacio ordinario puede tener dimensión superior a 3, ya que cada punto del espacio puede ser centro de una esfera de radio arbitrariamente pequeño, cuya superficie tiene dimensión 2. Pero, en la matemática moderna, se utiliza la palabra «espacio» para representar un sistema cualquiera de objetos para los cuales se ha definido el concepto de «distancia» y de «entorno» (véase Cap. VI, Sec. V, 4). Estos «espacios» abstractos pueden tener más de tres

[1] No pretende ser ésta una demostración rigurosa de que el plano tiene dimensión 2, de acuerdo con nuestra definición, ya que supone que la circunferencia de un círculo es de dimensión 1 y que el plano no tiene dimensión 0 ni 1. Pero puede darse una demostración de estos hechos y de sus análogos para dimensiones mayores. Esta demostración nos dice que la definición de dimensión de un conjunto general de puntos no contradice los resultados conocidos de los conjuntos más sencillos.

dimensiones. Un ejemplo sencillo es el *espacio cartesiano de* n *dimensiones*, cuyos «puntos» son conjuntos ordenados de n números reales:

$$P = (x_1, x_2, x_3, \ldots, x_n),$$
$$Q = (y_1, y_2, y_3, \ldots, y_n);$$

y en el cual se define la «distancia» entre dos puntos P y Q por la igualdad

$$d(P, Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

Puede demostrarse que este espacio tiene n dimensiones. Un espacio que no tiene dimensión n, cualquiera que sea el entero n, se dice que es de dimensión infinita. Se conocen muchos ejemplos de tales espacios.

Uno de los hechos más interesantes de la teoría de la dimensión es la siguiente propiedad característica de las figuras de dos, tres y, en general, de n dimensiones. Consideremos primero el caso bidimensional. Si se subdivide cualquier figura sencilla bidimensional en un número de regiones suficientemente pequeñas (cada una de las cuales incluya a su frontera), existirán necesariamente puntos, en los que se encuentran *tres o más* de estas regiones, *independientemente de su forma*. Además, *existen subdivisiones* de la figura para las que cada punto pertenece, *a lo sumo*, a tres regiones de la subdivisión. Así, si la figura bidimensional es un cuadrado, como en la figura 131, existe un punto que pertenece a las tres regiones 1, 2, 3; aunque para esta subdivisión particular, ningún punto pertenece a más de tres regiones. Análogamente, en el caso tridimensional puede demostrarse que, si se descompone un volumen en otros suficientemente pequeños, existen siempre puntos comunes por lo menos a cuatro de estos últimos, y que para una subdivisión adecuadamente elegida, no más de cuatro tendrán un punto común.

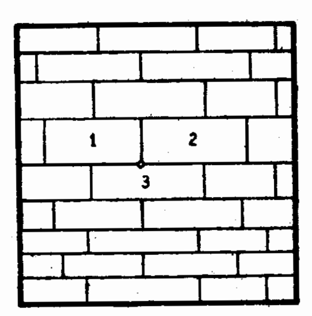


FIG. 131.—Teorema del recubrimiento.

Estas observaciones sugieren el siguiente teorema, debido a Lebesgue y Brouwer. Si se divide una figura de n dimensiones, de cualquier manera, en otras subregiones suficientemente pequeñas, siempre existirán puntos que pertenecerán, por lo menos, a $n + 1$ de ellas; además, en todos los casos es posible encontrar una subdivisión en regiones arbitrariamente pequeñas para la cual ningún punto pertenezca a más de $n + 1$ regiones. Debido a la forma en que se realiza la subdivisión, este teorema se llama también teorema del «recubrimiento». Caracteriza la dimensión de cualquier figura geométrica: aquellas figuras para las cuales el teorema es válido son n dimensionales, mientras todas las demás tendrán otra dimensión. Por esta razón, se puede tomar como *definición* de dimensión, tal como hacen algunos autores.

La dimensión de un conjunto cualquiera es una característica topológica del conjunto; dos figuras de diferente dimensión no pueden ser equivalentes topológicamente. Éste es el famoso teorema topológico de la «invariabilidad de la dimensión», cuya importancia resalta aún más, si se le compara con el hecho ex-

puesto en la página 94, referente a que el conjunto de puntos de un cuadrado tiene el mismo número cardinal que el de los puntos de un segmento rectilíneo. La correspondencia allí definida no es topológica, puesto que no se cumplen las condiciones de continuidad.

***4. Un teorema de punto invariante.**—En las aplicaciones de la topología a las otras ramas de la matemática, los teoremas referentes a «puntos fijos» desempeñan un importante papel. Un ejemplo típico es la siguiente proposición de Brouwer. Para la intuición resulta mucho menos evidente que la mayoría de los hechos topológicos.

Consideremos un disco circular en el plano, con lo cual queremos significar el interior de un círculo, junto con su circunferencia. Supongamos que se someten los puntos de ese disco a una transformación continua cualquiera (que no precisa ser biunívoca), en la cual cada punto permanece interior al círculo, aunque cambie su situación. Por ejemplo, puede encogerse, doblarse, estirarse o deformarse de cualquier manera un delgado disco de goma, en tanto que la posición final de cada punto del disco permanezca interior a la circunferencia original. Así también, si se pone en movimiento el líquido contenido en una vasija, agitándolo de tal manera que las partículas de la superficie permanezcan en ella, pero cambiando su posición, en un momento determinado la posición de esas partículas de la superficie define una transformación continua de su distribución original. El teorema de Brouwer afirma que *toda transformación de ese tipo deja invariable la posición de un punto por lo menos*, es decir, existe al menos un punto cuya posición, después de la transformación, coincide con la que tenía inicialmente. (En el ejemplo de la superficie del líquido, el punto fijo, en general, cambiará de posición con el tiempo, aunque para una simple rotación circular el centro permanece siempre inmóvil.) La demostración de la existencia de un punto fijo es típica de la clase de razonamiento utilizado para establecer muchos teoremas topológicos.

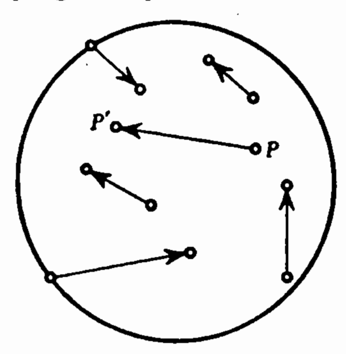


FIG. 132. — Vectores de transformación.

Consideremos el disco antes y después de la transformación y supongamos, contra lo que afirma el teorema, que *ningún* punto conserva su posición anterior, por lo que, en la transformación, cada punto se traslada a otro punto, situado dentro del círculo o en su

circunferencia. Asignemos a cada punto P del disco primitivo una flecha o «vector» en la dirección PP', siendo P' la imagen del punto P en la transformación. Para todo punto del disco existe dicho vector, pues se ha supuesto que cada uno de ellos variaba su posición. Consideremos ahora los puntos de la circunferencia del círculo con sus vectores asociados. Todos ellos están dirigidos hacia el interior del círculo, pues, por hipótesis, ningún punto se transforma en otro exterior al círculo. Comencemos por un punto cualquiera P_1 del contorno y recorramos la circunferencia en sentido contrario a las agujas de un

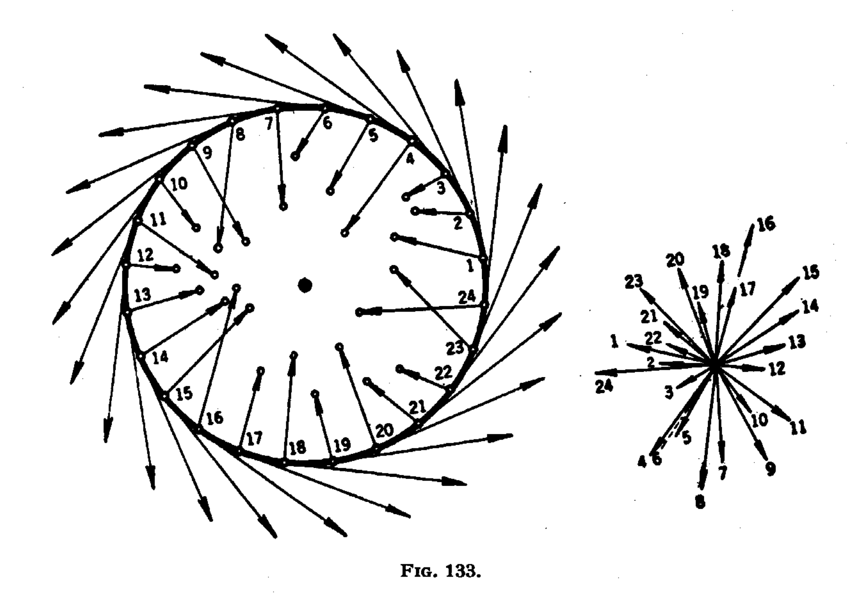


FIG. 133.

reloj. Al proceder así, cambiará la dirección del vector, pues los puntos del contorno tienen asociados vectores de dirección variable; las direcciones de estos vectores pueden ponerse de manifiesto trazando vectores paralelos por un punto del plano. Observemos que, al atravesar el círculo partiendo de P_1 para volver de nuevo a P_1, el vector gira y vuelve a su posición original. Designemos el número completo de vueltas dadas por este vector como el «índice» de los vectores del círculo; con más precisión, definamos el índice como la *suma algebraica* de las distintas variaciones angulares de los vectores, de forma que las efectuadas en el sentido de las agujas del reloj se tomarán con signo negativo, mientras que consideraremos como positivas las realizadas en sentido contrario. El índice es la suma final que, *a*

priori, puede ser cualquiera de los números 0, ± 1, ± 2, ± 3, ... , correspondiendo a una variación total del ángulo de 0°, ±360°, ±720°, ... Nuestra afirmación es que *el índice es igual a* 1; esto es, que la variación total en la dirección del vector es exactamente una vuelta positiva completa. Para probarlo, recordemos que el vector asociado a cualquier punto P de la circunferencia está dirigido siempre hacia el interior del círculo y nunca en la dirección de la tangente. Ahora bien: si este vector girase un ángulo total distinto del ángulo total girado por el vector *tangente* (que es de 360°, ya que el vector tangente efectúa evidentemente una revolución positiva completa), entonces la diferencia entre los ángulos totales girados por el vector tangente y el vector transformación sería un múltiplo (distinto de 0) de 360°, ya que efectúa un número entero de revoluciones. De ahí que el vector transformación deba girar completamente alrededor de la tangente por lo menos una vez, durante el circuito completo de P_1 a P_1; y puesto que el vector transformación y el tangente giran continuamente, en un cierto punto de la circunferencia el vector transformación debe coincidir con el tangente, lo que, como ya hemos visto, es imposible.

Si consideramos ahora una circunferencia cualquiera, concéntrica con la del disco e interior a ella, junto con los correspondientes vectores transformación de esta nueva circunferencia, el índice de dichos vectores para esta circunferencia debe ser 1. Pues si pasamos con continuidad de la circunferencia a cualquiera otra concéntrica, el índice debe variar continuamente, ya que las direcciones de los vectores transformación varían continuamente de punto a punto, dentro del disco. Pero como el índice puede tomar sólo valores enteros, permanecerá constantemente igual a 1, puesto que si saltara de 1 a cualquier otro valor entero habría una discontinuidad en el valor del índice. (La conclusión de que debe ser constante una cantidad que varía continuamente, pero que sólo puede tomar valores enteros, es una muestra típica de una clase de razonamiento matemático que interviene en muchas demostraciones.) Así podemos encontrar un círculo concéntrico tan pequeño como se quiera, para el cual el índice de los vectores transformación correspondientes es 1. Pero esto es imposible, puesto que de la supuesta continuidad de la transformación, los vectores de un círculo suficientemente pequeño tendrán aproximadamente la misma dirección que el vector del centro del círculo. Así, pues, el cambio total de sus ángulos puede llegar a ser tan pequeño como se quiera, p. ej., menor que 10°, tomando un círculo suficientemente pequeño. En consecuencia, como el índice debe ser

un número entero, será igual a cero. Esta contradicción demuestra que es falsa nuestra hipótesis inicial, según la cual no existía ningún punto fijo en la transformación, con lo que queda completada la demostración.

El teorema que acabamos de demostrar no sólo es válido para un disco, sino para cualquier otra región triangular o cuadrada, o para cualquier superficie que sea imagen de un disco en una transformación topológica. Pues si A es una figura cualquiera, correlacionada con un disco mediante una transformación biunívoca y continua, una transformación continua de A en sí misma, que no tuviera punto fijo, definiría una transformación continua del disco en sí mismo, sin punto fijo, y ya hemos demostrado que esto es imposible. El teorema conserva su validez en tres dimensiones, para esferas o cubos, pero la demostración no es tan sencilla.

Aunque para la intuición no resulta demasiado evidente el teorema del punto fijo de Brouwer, es fácil demostrarlo como consecuencia inmediata del hecho siguiente, cuya evidencia es mucho más intuitiva: es imposible transformar con continuidad un disco circular en su circunferencia exclusivamente, de tal modo que todo punto de ésta permanezca fijo. Demostraremos que la existencia de una transformación, sin punto fijo, de un disco en sí mismo contradice dicha proposición. Supongamos que $P \to P'$ fuera una transformación de ese tipo; para cada punto P del disco podemos trazar una flecha, que comienza en P' y continúa, pasando por P, hasta que alcanza la circunferencia en algún punto P^*. Entonces, la transformación $P \to P^*$ sería continua y tal que transformaría el disco completo en su circunferencia exclusivamente, dejando invariable todo punto de la circunferencia, contra la hipótesis de que una tal transformación es imposible. Puede utilizarse un razonamiento similar para establecer el teorema de Brouwer en tres dimensiones, para una esfera o un cubo.

Es fácil ver que algunas figuras geométricas admiten transformaciones en sí mismas, continuas y sin puntos fijos; p. ej., la región anular comprendida entre dos circunferencias concéntricas admite, como transformación continua sin puntos fijos, una rotación de ángulo cualquiera, no múltiplo de 360°, alrededor de su centro. La superficie esférica admite, como transformación continua sin puntos fijos, la que lleva cada punto al diametralmente opuesto. Pero puede demostrarse mediante razonamiento análogo al utilizado para el disco, que cualquier otra transformación continua que no haga corresponder a ningún punto el diametralmente opuesto (p. ej., cualquier pequeña deformación) tiene un punto fijo.

Los teoremas del punto fijo tales como éstos proporcionan un método potente para demostrar muchos «teoremas de existencia» que a primera vista no parecen tener carácter geométrico. Un ejemplo famoso es un teorema de punto fijo, sospechado por Poincaré en 1912, poco antes de su muerte. Este teorema tiene consecuencias inmediatas en la existencia de un número infinito de órbitas periódicas en el problema restringido de los tres cuerpos. Poincaré no pudo confirmar su sospecha, y constituyó una gran hazaña de los matemáticos americanos el que, al año siguiente, G. D. Birkhoff consiguiera dar una demostración. Desde

entonces, los métodos topológicos se aplican con gran éxito al estudio del comportamiento cualitativo de los sistemas dinámicos.

5. Nudos.—Como último ejemplo, indicaremos que el estudio de los nudos ofrece difíciles problemas matemáticos de carácter topológico. Un nudo se hace entrelazando un trozo de cuerda, después de lo cual se unen los extremos. La curva cerrada resultante representa una figura geométrica que sigue siendo esencialmente la misma después de deformar o retorcer la cuerda, sin romperla. Pero ¿cómo es posible dar una caracterización intrínseca que permita distinguir una curva cerrada, con nudos, situada en el espacio, de otra sin ellos, como, p. ej., una circunferencia? La respuesta no es en modo alguno sencilla y menos todavía lo es el análisis matemático completo de las

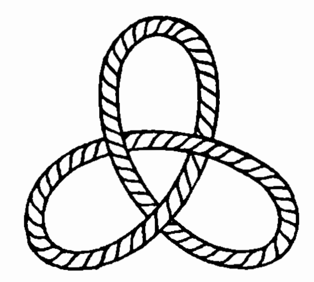
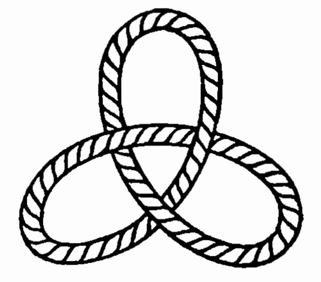



FIG. 134.—Nudos topológicamente equivalentes, que no se pueden deformar uno en el otro.

distintas clases de nudos y de sus diferencias mutuas. Aun para los casos más sencillos esto ha resultado una tarea ímproba. Considérense los dos nudos triples que aparecen en la figura 134. Ambos son completamente simétricos, «imágenes especulares» uno del otro; son topológicamente equivalentes, pero no congruentes. Surge el problema de saber si es posible deformar uno de estos nudos en el otro de una forma continua. La respuesta es negativa, pero la demostración de este hecho requiere mayores conocimientos de la técnica topológica y de la teoría de grupos que es posible dar aquí.

IV. CLASIFICACIÓN TOPOLÓGICA DE LAS SUPERFICIES

1. Género de una superficie.—Al estudiar las superficies de dos dimensiones se plantean muchos e importantes problemas topológicos; p. ej., comparemos la superficie de una esfera con la de un toro.

Resulta evidente, de la figura 135, que ambas superficies difieren de un modo fundamental. Sobre la esfera, como en el plano, toda curva simple cerrada, tal como C, divide a la superficie en dos partes, mientras que en el toro existen curvas cerradas, tales como C', que no

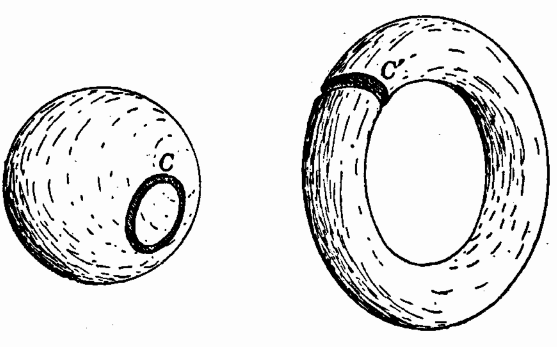


FIG. 135.—Cortes en la esfera y el toro.

dividen a la superficie en dos partes. Decir que C divide a la superficie esférica en dos partes significa que, si ésta se corta a lo largo de C, se obtendrán dos trozos distintos e inconexos, lo que equivale a poder encontrar dos puntos de la superficie esférica tales que cualquier curva esférica que los una deba cortar a C. Por otra parte, si se corta el toro a lo largo de la curva cerrada C', la superficie resultante se mantiene todavía unida; cualquier punto de la superficie puede unirse con otro de la misma, mediante una curva que no corta a C'.

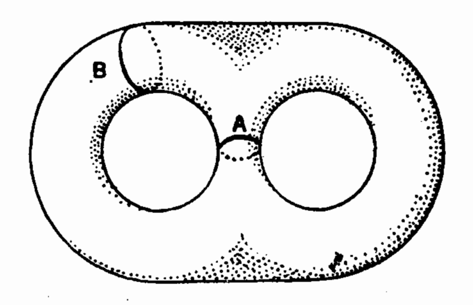


FIG. 136.—Superficie de género 2.

Esta diferencia entre la esfera y el toro nos dice que ambos tipos de superficies son topológicamente distintos y nos prueba que es imposible transformar una en la otra de una manera continua.

Consideremos ahora la superficie con dos agujeros representada en la figura 136. En ella se pueden trazar *dos* curvas cerradas, A y B, no mutuamente secantes, que no dividen a la superficie; en cambio, el toro queda dividido siempre en dos partes por dos cualesquiera de tales curvas. Por otra parte, *tres* curvas cerradas, que no se corten entre sí, dividen siempre a una superficie con dos agujeros.

Estos hechos nos inducen a definir el *género* de una superficie como el número máximo de curvas simples cerradas, no secantes entre sí, que pueden trazarse sobre la superficie sin dividirla. El género de la esfera es 0, el del toro es 1, mientras que el de la superficie de la figura 136 es 2. Una superficie similar con p agujeros tiene género p. El género es una propiedad topológica de una superficie que permanece invariable al deformarla. Recíprocamente, puede probarse (omitimos la demostración) que si dos superficies cerradas tienen el mismo género, es posible deformar una en la otra, por lo que el género $p = 0, 1, 2, \ldots$ de una superficie cerrada la caracteriza completamente desde el punto de vista topológico. (Suponemos que las superficies consideradas son superficies cerradas ordinarias de «dos caras». Más adelante consideraremos superficies de «una sola cara».) Por ejemplo, el buñuelo con dos agujeros y la esfera con dos *asas* de la figura 137 son ambas superficies cerradas de género 2; y es evidente que se puede deformar con continuidad cualquiera de las dos hasta obtener la otra. Puesto que el buñuelo con p agujeros, o su equivalente, la esfera de p asas, es de género p, podemos tomar cualquiera de ellas como representante topológico de todas las superficies cerradas de género p.

***2. Caracterización euleriana de una superficie.**—Supongamos que una superficie cerrada S de género p se divide en un cierto número de

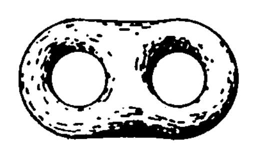
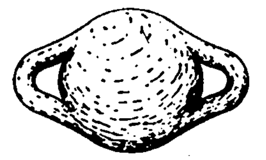
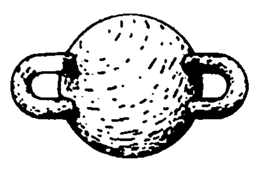




FIG. 137.—Superficies de género 2.

regiones, marcando vértices sobre S y uniéndolos por arcos de curva. Vamos a demostrar que

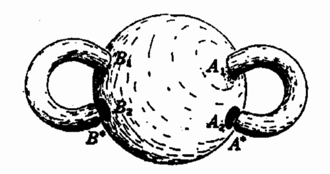
$$V - A + C = 2 - 2p, \qquad [1]$$

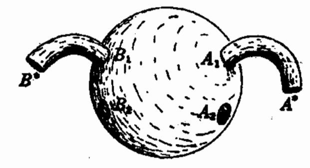
donde V = número de vértices, A = número de arcos y C = número de regiones. El número $2 - 2p$ se llama *característica de Euler* de la superficie. Ya hemos visto que para la esfera $V - A + C = 2$, lo que coincide con [1], pues en este caso, $p = 0$.

Para demostrar la fórmula general [1] podemos suponer que S es una esfera de p asas, pues, como ya hemos dicho, cualquier superficie de género p puede deformarse continuamente en tal superficie, y

durante esta deformación permanecerán invariables los números $V - A + C$ y $2 - 2p$. Elegiremos la deformación de tal manera que las curvas cerradas A_1, A_2, B_1, B_2 ..., por donde las asas se unen a la esfera, estén formadas por arcos de la subdivisión dada (véase figura 138, que aclara la demostración para el caso $p = 2$).

Cortemos ahora la superficie S a lo largo de las curvas A_2, B_2... y enderecemos las asas. Cada una tendrá una arista libre, limitada por una nueva curva A^*, B^*, ... con el mismo número de vértices y arcos que A_2, B_2 ..., respectivamente. Por tanto, $V - A + C$ no variará, puesto que los nuevos vértices adicionales compensarán exactamente





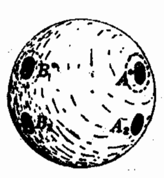


FIG. 138.

los nuevos arcos, en tanto que no se crean así nuevas regiones. Deformemos ahora la superficie achatando las asas salientes, hasta que la superficie resultante sea simplemente una esfera, de la cual se han suprimido $2p$ regiones. Como $V - A + C$ es igual a 2 para cualquier subdivisión de la esfera completa, tendremos

$$V - A + C = 2 - 2p$$

para la esfera de la que se han suprimido $2p$ regiones, o sea, también para la primitiva esfera de p asas, según queríamos demostrar.

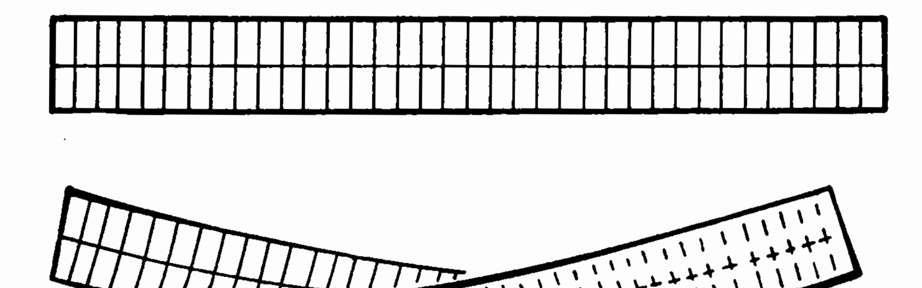
La figura 121 aclara la aplicación de la fórmula [1] para una superficie S compuesta por polígonos planos. Dicha superficie puede deformarse continuamente hasta convertirla en un toro, de tal forma que el género p sea igual a 1 y se tenga: $2 - 2p = 2 - 2 = 0$. Como prevé la fórmula [1]

$$V - A + C = 16 - 32 + 16 = 0.$$

Ejercicio: Subdivídase en regiones el buñuelo con dos agujeros de la figura 137 y demuéstrese que $V - A + C = -2$.

3. Superficies uniláteras.—Una superficie ordinaria tiene siempre dos caras, lo cual se aplica tanto a las superficies cerradas (la esfera o el toro) como a las superficies limitadas por curvas (el disco, o un toro del cual se ha suprimido un pedazo). Las dos caras de una su-

perficie de esa clase podrían pintarse de distintos colores para distinguirlas. Si la superficie es cerrada, los dos colores quedan separados; si está limitada por curvas, los dos colores se encuentran sólo a lo largo de éstas. Un insecto que caminase por una de esas superficies,



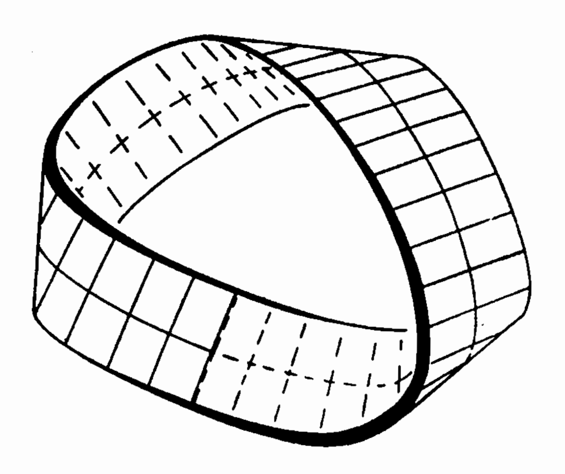


FIG. 139.—Construcción de la cinta de Moebius.

y al que se le impidiera atravesar las curvas que forman el contorno (si existe), permanecería siempre sobre la misma cara.

Moebius hizo el descubrimiento sorprendente de que existen superficies de una *sola* cara. La más sencilla de dichas superficies, llamada cinta o banda de Moebius, se forma tomando una tira larga y rectangular de papel y pegando sus dos extremos después de darle media vuelta, como indica la figura 139. Un insecto que recorriera

esta superficie, manteniéndose siempre en el eje de la cinta, volvería a su posición de partida.

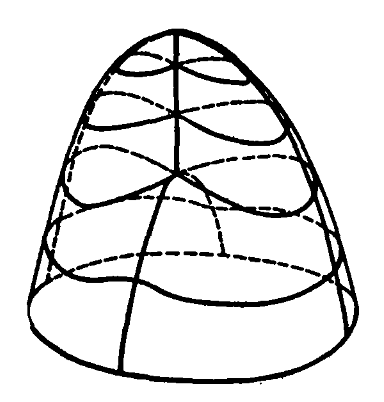


FIG. 140.—Cofia cruzada.

La banda de Moebius tiene sólo un borde, pues su frontera consiste en una curva cerrada única. La superficie ordinaria de dos caras, formada pegando los dos extremos de un rectángulo, sin retorcerlo, está limitada por dos curvas distintas. Si se corta esta última banda, a lo largo de la línea central, se descompone en otras dos bandas del mismo tipo. Pero si se corta la superficie de Moebius a lo largo de dicha línea (Fig. 139), vemos que sigue siendo de una sola pieza. No es fácil que quien no conozca previamente la superficie de Moebius prevea esto, pues es opuesto a lo que la intuición nos dice que «debiera» ocurrir. Si la superficie que resulta después de cortar la cinta de Moebius a lo largo de su eje se corta de nuevo en igual forma, resultan dos nuevas superficies del mismo tipo, pero entrelazadas.

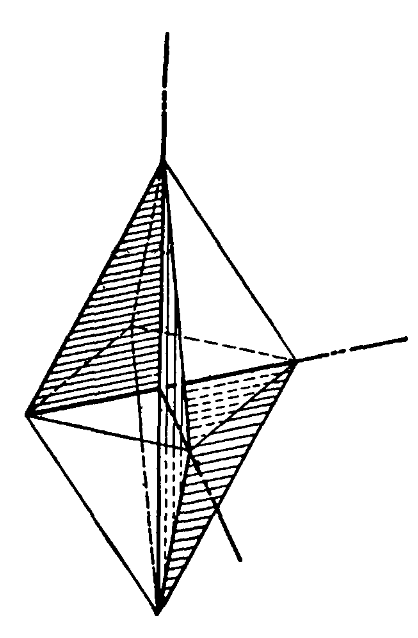


FIG. 141.—Cinta de Moebius de contorno plano triangular.

Constituye un pasatiempo entretenido cortar dichas superficies a lo largo de líneas paralelas al borde a una distancia de 1/2, 1/3, etc., de la anchura total. El contorno de la cinta de Moebius es una curva cerrada, simple y sin nudos, que puede convertirse por deformación en una curva plana; p. ej., una circunferencia. Al efectuar la deformación puede permitirse que la superficie se corte a sí misma, de modo que resulte una superficie de una sola cara que se corta a sí misma, como está representada en la figura 140 (cofia cruzada). Se considera que la línea de autointersección está compuesta en realidad de otras dos, cada una

de ellas perteneciente a una de las dos porciones que se cortan en ella. La unilateralidad de la cinta de Moebius queda conservada, puesto que dicha propiedad es topológica; es imposible deformar continuamente una superficie de una sola cara en otra de dos caras. Aunque resulta sorprendente, es posible realizar la deformación de manera que el contorno de la cinta de Moebius se haga plano (p. ej., triangular), mientras la banda permanece sin cortarse a sí misma. La figura 141 representa dicho modelo, que ha sido ideado por B. Tuckermann. El contorno es un triángulo, mitad de un cuadrado diagonal de un octaedro regular. La banda consta de seis caras del octaedro y cuatro triángulos rectángulos, cada uno un cuarto de un plano diagonal.

Otra superficie unilátera interesante es la *botella de Klein.* Es una superficie cerrada, pero no tiene ni exterior ni interior. Topológicamente equivale a un par de cofias cruzadas, cuyos contornos coinciden.

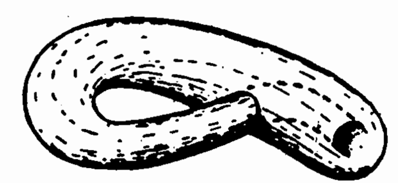


FIG. 142.—Botella de Klein.

Puede demostrarse que cualquier superficie cerrada, de *una sola cara,* de género $p = 1, 2, \ldots$, es topológicamente equivalente a una esfera, en la cual se han suprimido p discos y se los ha reemplazado por otras tantas cofias cruzadas. De ahí se deduce fácilmente que la característica euleriana $V - A + C$ de una superficie de ese tipo está relacionada con p por la ecuación

$$V - A + C = 2 - p.$$

La demostración es análoga a la correspondiente para superficies de dos caras. Demostraremos, primero, que la característica euleriana de una cofia cruzada, o de una cinta de Moebius, es 0. Para conseguirlo, observaremos que si se corta de través una superficie de Moebius, subdividida previamente en un cierto número de regiones, se obtiene un rectángulo que contiene dos vértices más, una arista más y el mismo número de regiones que la superficie de Moebius. Para el rectángulo, $V - A + C = 1$, como se ha visto en la página 251. De ahí que, para la superficie de Moebius, sea $V - A + C = 0$. El lector puede completar la demostración como ejercicio.

Resulta mucho más sencillo estudiar la naturaleza topológica de superficies como éstas mediante polígonos planos, en los cuales se identifican conceptualmente ciertos pares de aristas (véase Cap. IV, Apéndice). En los diagramas de la figura 143 las flechas se hacen coincidir, real o conceptualmente, en posición y dirección.

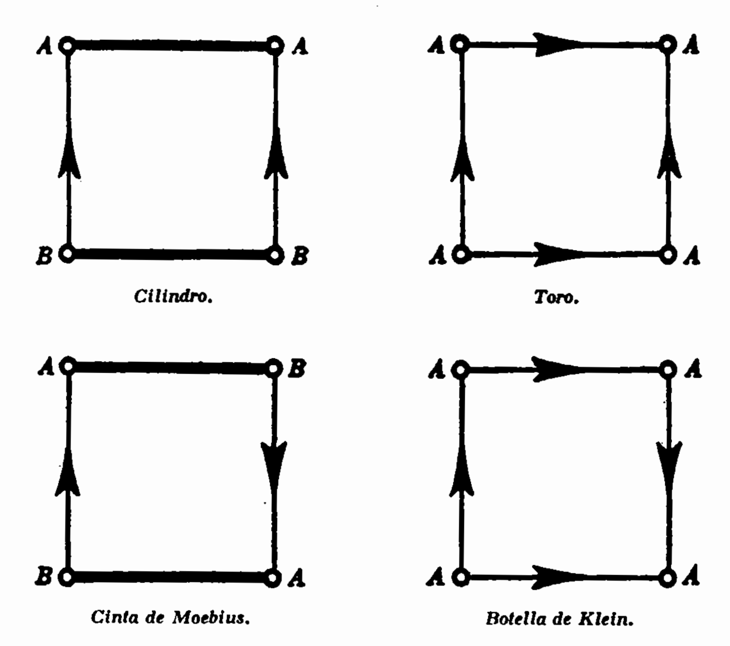


FIG. 143.—Superficies cerradas definidas por coordinación de aristas en una figura plana.

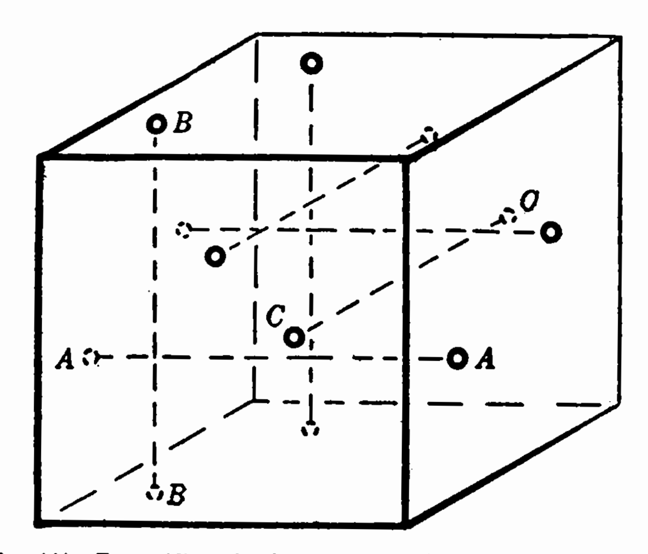


FIG. 144.—Toro tridimensional definido por identificación de contorno.

Puede utilizarse también este método de identificación para definir variedades cerradas tridimensionales, análogas a las superficies cerradas bidimensionales; p. ej., si identificamos los puntos correspondientes de las caras opuestas de un cubo (Fig. 144), obtenemos una variedad cerrada tridimensional, llamada toro tridimensional. Esta variedad equivale topológicamente al espacio comprendido entre dos toros

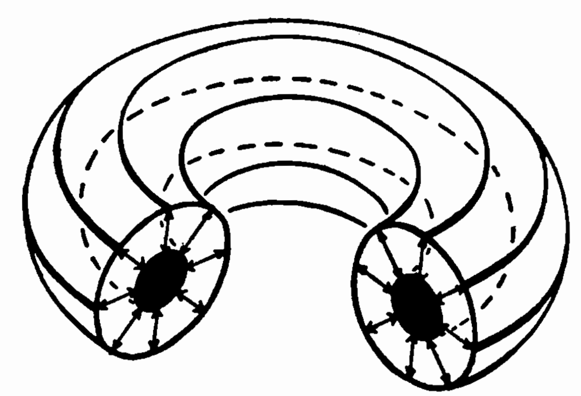


FIG. 145.—Otra representación del toro tridimensional (la sección indica la identificación).

concéntricos, uno dentro del otro, en el cual se identifican los puntos correspondientes de ambas superficies (Fig. 145). Esta última variedad resulta a partir de un cubo en el cual se hagan coincidir dos pares de caras identificadas conceptualmente.

APÉNDICE

***1. El teorema de los cinco colores.**—Basándonos en la fórmula de Euler podemos demostrar que es posible iluminar adecuadamente cualquier mapa sobre una superficie esférica utilizando, a lo más, cinco colores distintos. (Según lo dicho en la pág. 258, se considera que un mapa está adecuadamente iluminado si dos regiones contiguas, que tienen como límite común todo un segmento de sus fronteras, no están pintadas con el mismo color.) Nos limitaremos a aquellos mapas cuyas regiones están limitadas por polígonos simples cerrados, compuestos de arcos circulares. Podemos suponer también que en cada vértice se encuentran exactamente tres arcos; diremos que un mapa de este tipo es *regular*. Si reemplazamos cada vértice en el que concurran más de tres arcos por un pequeño círculo y unimos el interior

de cada uno de esos círculos a una de las regiones que se encuentran en el vértice, tenemos un nuevo mapa, en el cual los vértices múltiples vienen reemplazados por un cierto número de vértices triples. El nuevo mapa contendrá el mismo número de regiones que el primitivo. Si este nuevo mapa, que es regular, se puede iluminar adecuadamente con cinco colores, haciendo que disminuya el radio de los círculos hasta convertirse en puntos, tendremos iluminado en la forma deseada el mapa primitivo. Por ello, basta demostrar que cualquier mapa regular sobre la superficie esférica puede iluminarse con cinco colores.

En primer lugar, demostraremos que todo mapa regular debe contener al menos un polígono de menos de seis lados. Sea F_n el número de regiones de n lados en un mapa regular; si es C el número total de regiones,

$$C = F_2 + F_3 + F_4 + \cdots \qquad [1]$$

Cada arco tiene dos extremos y en cada vértice se encuentran tres arcos. De ahí que si A indica el número de arcos del mapa y V el de vértices,

$$2A = 3V. \qquad [2]$$

Además, una región limitada por n arcos tiene n vértices, y cada vértice pertenece a tres regiones, por lo que

$$2A = 3V = 2F_2 + 3F_3 + 4F_4 + \cdots. \qquad [3]$$

Por la fórmula de Euler, se tiene

$$V - A + C = 2, \quad \text{o} \quad 6V - 6A + 6C = 12.$$

De [2] se deduce que $6V = 4A$, por lo que $6C - 2A = 12$.

Por tanto, de [1] y [3] resulta:

$$6(F_2 + F_3 + F_4 + \cdots) - (2F_2 + 3F_3 + 4F_4 + \cdots) = 12,$$

o

$$(6 - 2)F_2 + (6 - 3)F_3 + (6 - 4)F_4 + (6 - 5)F_5 + (6 - 6)F_6 + \\ + (6 - 7)F_7 + \cdots = 12.$$

Por lo menos uno de los términos del primer miembro debe ser positivo, así que, al menos, uno de los números F_2, F_3, F_4, F_5 es positivo, como queríamos demostrar.

Vamos a probar ahora el teorema de los cinco colores. Sea M un mapa regular cualquiera sobre la superficie esférica, que tiene en total n regiones. Sabemos que por lo menos una de estas regiones debe tener menos de seis lados.

Caso 1. M contiene una región A de 2, 3 ó 4 lados. En este caso, se suprime la frontera entre A y una de las regiones limítrofes. (Si A tiene cuatro lados, una región puede extenderse y tocar dos lados de A no adyacentes. En este caso, por el teorema de la curva de Jordan, quedarán diferenciadas las regiones que tocan los otros dos lados de A, y suprimimos la frontera entre A y una de las últimas regiones.)

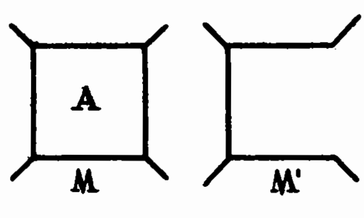


FIG. 146.

El mapa resultante, M', será regular, con $n - 1$ regiones. *Si M' se puede iluminar con cinco colores, lo mismo puede hacerse con M.* Dado que a lo más cuatro regiones de M limitan con A, siempre podremos encontrar un quinto color para A.

Caso 2. M contiene una región A con cinco lados. Consideremos las cinco regiones limítrofes con A y llamémoslas B, C, D, E, y F. Podremos encontrar siempre entre ellas un par sin frontera común; pues si, p. ej., B y D se tocan, impiden que C toque a E o F, ya que

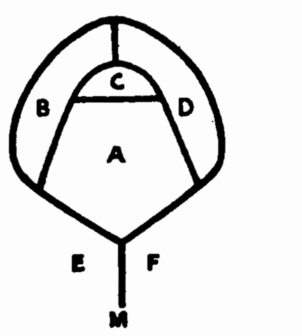
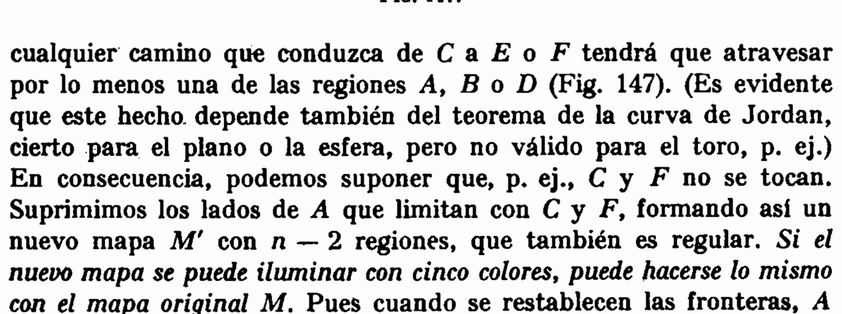


FIG. 147.


cualquier camino que conduzca de C a E o F tendrá que atravesar por lo menos una de las regiones A, B o D (Fig. 147). (Es evidente que este hecho depende también del teorema de la curva de Jordan, cierto para el plano o la esfera, pero no válido para el toro, p. ej.) En consecuencia, podemos suponer que, p. ej., C y F no se tocan. Suprimimos los lados de A que limitan con C y F, formando así un nuevo mapa M' con $n - 2$ regiones, que también es regular. *Si el nuevo mapa se puede iluminar con cinco colores, puede hacerse lo mismo con el mapa original M.* Pues cuando se restablecen las fronteras, A

no estará en contacto con más de cuatro colores diferentes, ya que C y F tienen el mismo color, y siempre podremos encontrar un quinto color para A.

Así, en cualquiera de los casos, si M es un mapa regular de n regiones, podemos construir otro nuevo mapa regular, M', que tiene $n-1$ ó $n-2$ regiones, y tal que si se puede iluminar M' con cinco colores, cabe hacer lo mismo con M. Este proceso se puede aplicar también a M', etc., lo que conduce a una sucesión de mapas deducidos de M:

$$M, M', M'', \cdots$$

Dado que el número de regiones de esta sucesión de mapas decrece constantemente, debemos llegar al fin a un mapa con no más de cinco regiones, el cual se puede siempre iluminar con cinco colores a lo sumo. De ahí, volviendo paso a paso hasta M, vemos que también éste puede iluminarse con cinco colores, con lo que queda terminada la demostración. Obsérvese que la prueba es constructiva, ya que proporciona un método perfectamente practicable, aunque laborioso, para iluminar efectivamente cualquier mapa de n regiones en un número finito de pasos.

2. El teorema de la curva de Jordan para polígonos.—El teorema de Jordan dice que cualquier curva simple y cerrada C divide a los puntos del plano no situados sobre C en dos recintos distintos (que no tienen punto alguno común), de los cuales C es la frontera común. Daremos una demostración de este teorema para el caso en que C sea un *polígono* cerrado P.

Probaremos que los puntos del plano que no se hallan sobre P quedan subdivididos en dos clases, A y B, de tal manera que dos puntos cualesquiera de la misma clase pueden unirse mediante una poligonal que no corta a P, mientras que cualquier poligonal que una un punto de A con otro de B debe cortar a P. La clase A formará el «exterior» del polígono, mientras que la B constituye el «interior».

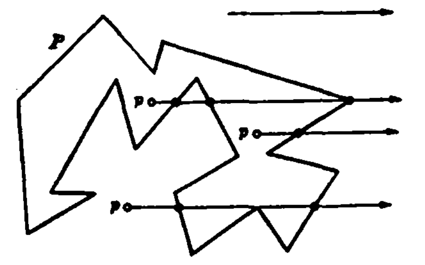


FIG. 148.—Cómputo de intersecciones.

Iniciaremos la demostración eligiendo una dirección fija en el plano, no paralela a ninguno de los lados de P; como P tiene un número finito de lados, esto es siempre posible. Definiremos ahora las clases A y B en la forma siguiente:

El punto p pertenece a A si el rayo que pasa por p en la dirección prefijada corta a P en un número *par* de puntos 0, 2, 4, 6, ... El punto p pertenece a B si el rayo que pasa por p en dicha dirección corta a P en un número *impar* de puntos 1, 3, 5, ...

En lo que respecta a las rayos que pasan por los vértices de P, no los contaremos como intersecciones si los lados de P que se cortan en ellos se encuentran del mismo lado del rayo; pero consideraremos que hay intersección si ambos lados se encuentran en distinta parte del rayo. Diremos que dos puntos, p y q, tienen la misma «paridad» si pertenecen a la misma clase, A o B.

Observaremos, en primer lugar, que todos los puntos de un segmento rectilíneo que no corta a P tienen la misma paridad. Pues la paridad de un punto p que se mueve a lo largo de un tal segmento, puede cambiar sólo cuando el rayo en la dirección fija trazado por p pasa por un vértice de P; y en ninguno de los casos posibles cambiará la paridad, debido al convenio que acabamos de hacer. De esto se deduce que *si se une un punto* p_1 *de* A *con otro* p_2 *de* B *mediante una poligonal, ésta debe cortar a* P; pues en otro caso, la paridad de todos los puntos de la trayectoria, y en particular de p_1 y p_2, sería la misma. Además, podemos demostrar que *dos puntos cualesquiera de la misma clase,* A *o* B, *pueden unirse mediante una poligonal que no corta a* P. Sean p y q los dos puntos; si el segmento rectilíneo pq, que une p con q, no corta a P, es la poligonal buscada. En caso contrario, sea p' el primer punto de intersección de ese segmento con P, y q' el último (Fig. 149). Construyamos la poligonal que arranca de p, a lo largo del segmento pp', que se desvía antes de llegar a p', y sigue a lo largo de P hasta que P encuentra a pq en q'. Si podemos demostrar que esta poligonal corta a pq entre q' y q y no entre p' y q', podrá continuarse aquélla hacia q, siguiendo $q'q$, sin cortar a P. Es evidente que dos puntos cualesquiera, r y s, muy próximos entre sí, pero a distinto lado de algún segmento de P, deben tener distinta paridad, pues el rayo que pasa por r cortará a P en un punto más que el rayo que pasa por s. Vemos así que la paridad cambia cuando atravesamos el punto q' a lo largo del segmento pq. Se sigue de ello que la poligonal de trazos debe cortar a pq entre q' y q, ya que

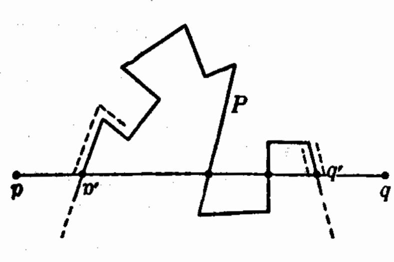


FIG. 149.

p y q (y, en consecuencia, cualquier otro punto de la línea de trazos) tienen la misma paridad.

Esto completa la demostración del teorema de la curva de Jordan para el caso de un polígono P. Puede identificarse ahora el «exterior» de P con la clase A, puesto que si nos alejamos lo suficiente a lo largo de un rayo cualquiera en la dirección fijada, llegaremos a un punto más allá del cual no existirá intersección con P, por lo que todos esos puntos tienen paridad 0, y pertenecen a A. Esto identifica el «interior» de P con la clase B. No importa lo complicado que sea el polígono P, pues siempre podremos determinar si un punto dado p del plano está dentro o fuera de P, trazando un rayo y contando el número de sus intersecciones con P. Si éste es impar, el punto p se encuentra dentro de P, y no puede salir de su interior sin atravesar P en algún punto. Si el número de puntos de intersección es par, el punto p es exterior a P (véase Fig. 128).

*Se puede demostrar el teorema de la curva de Jordan para los polígonos de la siguiente manera: defínase el *orden* de un punto p_0 respecto a una curva cerrada cualquiera C, que no pasa por p_0, como el número de revoluciones completas que efectúa el radio vector que une p_0 con un punto móvil p' sobre la curva, mientras éste recorre toda la curva una sola vez. Sea

A = el conjunto de los puntos p_0 que no se encuentran sobre P y que tienen un orden *par* respecto a P,

B = el conjunto de los puntos p_0 que no se encuentran sobre P y que tienen un orden *impar* respecto a P.

Entonces, los conjuntos A y B así definidos forman, respectivamente, el exterior y el interior de P. Se deja como ejercicio al lector efectuar la demostración con detalle.

****3. El teorema fundamental del álgebra.**—El «teorema fundamental del álgebra» dice que si

$$f(z) = z^n + a_{n-1}z^{n-1} + a_{n-2}z^{n-2} + \cdots + a_1 z + a_0 \qquad [1]$$

es un polinomio, en el cual $n \geqslant 1$ y $a_{n-1}, a_{n-2}, \ldots, a_0$, números complejos cualesquiera, existe un número complejo α tal que $f(\alpha) = 0$. Dicho de otra manera: *en el campo de los números complejos, toda ecuación polinómica tiene una raíz.* [Vimos en la pág. 110 que de ello se deduce la posibilidad de descomponer $f(z)$ en el producto de n factores lineales:

$$f(z) = (z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n),$$

siendo $\alpha_1, \alpha_2, \ldots, \alpha_n$, los *ceros* de $f(z)$.] Es notable que este teorema pueda demostrarse mediante consideraciones de carácter topológico, relacionadas con las que se utilizaron para demostrar el teorema de Brouwer del punto fijo.

Recordará el lector que un número complejo es un símbolo $x+iy$,

en el cual, tanto x como y son números reales e i tiene la propiedad de que $i^2 = -1$. Puede representarse el número complejo $x + iy$ mediante el punto del plano cuyas coordenadas, respecto a un par de ejes perpendiculares, son x e y. Si en ese plano introducimos coordenadas polares, tomando el origen y la dirección positiva del eje de las x como polo y eje polar respectivamente, podemos escribir:

$$z = x + iy = r\,(\cos \theta + i \operatorname{sen} \theta),$$

siendo $r = \sqrt{x^2 + y^2}$. Se deduce de la fórmula de De Moivre que

$$z^n = r^n\,(\cos n\theta + i \operatorname{sen} n\theta)$$

(véase pág. 105). Así, pues, si hacemos que el número complejo z describa una circunferencia de radio r y centro en el origen, z^n describirá n veces una circunferencia completa de radio r^n mientras z describe la suya una sola vez. Recordemos también que r, módulo de z,

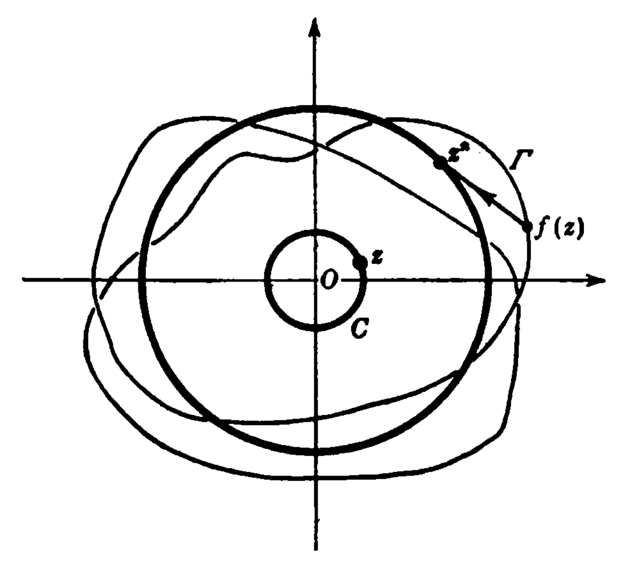


FIG. 150.—Demostración del teorema fundamental del álgebra.

que se expresa mediante el símbolo $|z|$, da la distancia de z a O, y que si $z' = x' + iy'$, la distancia entre z y z' es $|z - z'|$. Después de estos preliminares, podemos proceder a la demostración del teorema.

Supongamos que el polinomio [1] no tiene raíces, de donde resulta que para todo número complejo z se tiene:

$$f(z) \neq 0.$$

Supuesto esto, si hacemos que z describa cualquier curva cerrada en el plano x, y, $f(z)$ describirá una curva cerrada Γ, que no pasará por el origen (Fig. 150). Podemos, por tanto, definir el *orden* del origen O respecto a la función $f(z)$ para cualquier curva cerrada C como el *número total de revoluciones completas que efectúa el radio vector que une* O *con un punto de la curva* Γ *trazada por el punto representativo de* f(z), *mientras* z *recorre la curva C.* Como curva C tomaremos una circunferencia de centro O y de radio t y definimos la función $\varphi(t)$ como el orden de O respecto a la función $f(z)$ para el círculo de centro O y radio t. Es evidente que $\varphi(0) = 0$, puesto que un círculo de radio igual a cero es un punto y la curva Γ se reduce al punto $f(0) \neq O$. En el próximo párrafo demostraremos que $\varphi(t) = n$ para valores grandes de t. Pero el orden $\varphi(t)$ depende con continuidad de t, ya que $f(z)$ es una función continua de z. De ahí resulta una contradicción, pues la función $\varphi(t)$ sólo puede tomar valores enteros y, en consecuencia, no puede pasar con continuidad del valor 0 al valor n.

Queda sólo por probar que $\varphi(t) = n$ para valores grandes de t. Obsérvese que en un círculo de radio $z = t$, tal que

$$t > 1 \quad \text{y} \quad t > |a_0| + |a_1| + \cdots + |a_{n-1}|,$$

tenemos la desigualdad

$$\begin{aligned} |f(z) - z^n| &= |a_{n-1}z^{n-1} + \cdots + a_0| \leqslant |a_{n-1}| \cdot |z|^{n-1} + \\ &+ |a_{n-2}| \cdot |z|^{n-2} + \cdots + |a_0| = t^{n-1}\left[|a_{n-1}| + \cdots + \frac{|a_0|}{t^{n-1}}\right] \leqslant \\ &\leqslant t^{n-1}[|a_{n-1}| + |a_{n-2}| + \cdots + |a_0|] < t^n = |z^n|. \end{aligned}$$

Dado que el primer miembro es la distancia entre los puntos z^n y $f(z)$, mientras el último es la distancia del punto z^n al origen, vemos que el segmento rectilíneo que une los dos puntos $f(z)$ y z^n no puede pasar por el origen en tanto z se encuentre sobre la circunferencia de radio t y centro en el origen. Siendo esto así, podemos deformar continuamente la curva trazada por $f(z)$ hasta obtener la descrita por z^n, sin pasar nunca por el origen, trasladando simplemente todo punto de $f(z)$ a lo largo del segmento que lo une con z^n. Como el orden del origen varía con continuidad y puede tomar sólo valores enteros durante esa deformación, debe ser el mismo para ambas curvas. Como el orden de z^n es n, el de $f(z)$ ha de ser también n, con lo que la demostración queda ultimada.

CAPÍTULO VI

FUNCIONES Y LÍMITES

Introducción.—La parte principal de la matemática moderna se centra en torno a los conceptos de función y límite. En este capítulo los analizaremos de forma sistemática.

Una expresión tal como

$$x^2 + 2x - 3$$

no tiene valor definido mientras no se le asigne un valor a x; decimos que el valor de esta expresión es una *función* del valor de x, y escribimos:

$$x^2 + 2x - 3 = f(x).$$

Por ejemplo, para $x = 2$, resulta $2^2 + 2 \cdot 2 - 3 = 5$, por lo que $f(2) = 5$. De la misma manera, por sustitución directa, podemos hallar el valor de $f(x)$ para cualquier valor entero, fraccionario, irracional o incluso complejo de x.

El número de primos menores que n es una función $\pi(n)$ del entero n. Dado un valor de n, $\pi(n)$ está determinado, aunque no se conoce ninguna expresión algebraica que permita calcularlo. El área de un triángulo es función de las longitudes de sus tres lados; varía al variar éstas y está determinada cuando estas longitudes toman valores definidos. Si se somete un plano a una transformación proyectiva o topológica, las coordenadas de un punto, después de efectuada la transformación, dependen, es decir, son funciones de las coordenadas primitivas. El concepto de función aparece en cuanto se relacionan cantidades mediante una relación física determinada. El volumen de un gas encerrado en un cilindro es función de la temperatura y de la presión que ejerce el pistón. La presión atmosférica, observada en un globo, es función de su altitud sobre el nivel del mar. Todo el campo de los fenómenos periódicos—las mareas, las vibraciones de una cuerda, la emisión de ondas luminosas por un filamento incandescente—está regido por las funciones trigonométricas elementales sen x y cos x.

Para Leibniz (1646-1716), el primero que utilizó la palabra «función», y para los matemáticos del siglo XVIII, el concepto de relación funcional estaba más o menos identificado con la existencia de una fórmula matemática sencilla que expresara la naturaleza exacta de

esa relación. Este concepto resultó ser demasiado restrictivo para las necesidades de la física matemática, por lo que la idea de función, junto con el concepto anejo de límite, hubo de pasar por un largo proceso de generalización y clarificación, del cual daremos una sucinta exposición en este capítulo.

I. VARIABLE Y FUNCIÓN

1. Definiciones y ejemplos.—Se nos ofrecen a menudo entes matemáticos que estamos en libertad de elegir arbitrariamente entre todo un conjunto S de objetos. Llamaremos *variable* a un objeto de esa clase, perteneciente al *campo* o *dominio* S. Es costumbre utilizar las últimas letras del abecedario para designar las variables; p. ej., si S designa el conjunto de todos los números enteros, la variable X, definida en el dominio S, representa un entero arbitrario. Decimos que «la variable X varía en el conjunto S» significando que podemos identificarla con cualquier lemento del conjunto S. Conviene utilizar variables cuando deseamos hacer afirmaciones acerca de objetos elegidos arbitrariamente dentro de un conjunto; p. ej., si S sigue representando el conjunto de los números enteros, y X e Y son dos variables en el dominio S, la afirmación

$$X + Y = Y + X$$

es una expresión simbólica conveniente del hecho de que la suma de dos enteros es independiente del orden en que se tomen. Un caso particular está expresado por la igualdad

$$2 + 3 = 3 + 2,$$

en la cual aparecen sólo constantes. Para expresar la ley general, válida para cualquier par de números, se necesitan símbolos con el significado de variables.

No es necesario que el dominio S de la variable X sea un conjunto de números; p. ej., S puede ser el conjunto de todos los círculos del plano; X representará entonces un círculo determinado. O bien, S puede representar el conjunto de todos los polígonos cerrados del plano, siendo entonces X uno cualquiera de ellos. Tampoco es necesario que el dominio de una variable contenga un número infinito de elementos; p. ej., X puede denotar uno cualquiera de los miembros de la población S de una ciudad determinada en un momento prefijado. O bien, X puede significar uno cualquiera de los restos posibles cuando se divide un entero por 5; en este caso, el dominio S se compone de los cinco números 0, 1, 2, 3, 4.

El caso más importante de una variable numérica—para designarla se suele utilizar la letra x—es aquel en que el dominio de variabilidad S es un intervalo $a \leqslant x \leqslant b$ del eje numérico real. En este caso, diremos que x es una *variable continua* en dicho intervalo. El dominio de variabilidad de una variable continua puede extenderse hasta el infinito. Así, S puede ser el conjunto de todos los números reales y positivos, $x > 0$, o el conjunto de todos los números reales sin excepción. De manera similar, podemos considerar una variable X cuyos valores sean los puntos de un plano, o de un dominio dado del mismo, tal como el interior de un rectángulo o de un círculo. Puesto que cada punto del plano está definido por sus dos coordenadas x e y respecto a un par fijo de ejes, en este caso diremos usualmente que tenemos un *par de variables continuas*, x e y.

Puede ocurrir que a cada valor de la variable X esté asociado otro valor determinado de otra variable U. Se dice entonces que U es una *función* de X. El modo como ambas están enlazadas se expresa mediante un símbolo, tal como

$$U = F(X) \qquad \text{(léase: «}F\text{ de }X\text{»).}$$

Si X varía en el conjunto S, la variable U variará en otro conjunto, que llamaremos, p. ej., T. Si S representa el conjunto de todos los triángulos X del plano, se puede definir una función $F(X)$ asignando a cada triángulo la longitud $U = F(X)$ de su perímetro; T será entonces el conjunto de todos los números positivos. Observemos que dos triángulos distintos, X_1 y X_2, pueden tener el mismo perímetro, por lo que es posible la igualdad $F(X_1) = F(X_2)$, aunque sea $X_1 \neq X_2$. Una transformación proyectiva de un plano S en otro T asigna a cada punto X de S un solo punto U de T, de acuerdo con una regla perfectamente establecida, que podemos expresar mediante el símbolo funcional $U = F(X)$. En este caso, $F(X_1) \neq F(X_2)$ si $X_1 \neq X_2$, por lo que diremos que la representación de S en T es *biunívoca* (véase página 86).

A menudo, se definen las funciones de una variable continua mediante expresiones algebraicas; he aquí algunos ejemplos:

$$u = x^2, \qquad u = \frac{1}{x}, \qquad u = \frac{1}{1 + x^2}.$$

En la primera y en la última de estas expresiones, x varía en el conjunto completo de los números reales, mientras que en la segunda, x puede tomar cualquier valor real, excepto el 0, excluyéndose este valor, ya que $1/0$ no es un número.

El número $B(n)$ de los divisores primos de n es una función de n, en la cual n varía en el dominio de todos los números naturales. Con mayor generalidad, cualquier sucesión de números $a_1, a_2, a_3, \ldots$, puede considerarse como el conjunto de valores de una función $u = F(n)$, donde el dominio de variabilidad de la variable independiente n es el conjunto de los números naturales. Sólo por brevedad, designamos mediante a_n el n-ésimo término de la sucesión, en lugar de usar la notación funcional más explícita $F(n)$. Las expresiones consideradas en el capítulo primero:

$$S_1(n) = 1 + 2 + \cdots + n = \frac{n(n+1)}{2},$$

$$S_2(n) = 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6},$$

$$S_3(n) = 1^3 + 2^3 + \cdots + n^3 = \frac{n^2(n+1)^2}{4},$$

son funciones de la variable entera n.

Si $U = F(X)$, se reserva de ordinario el nombre de *variable independiente* para X, mientras que U se llama *variable dependiente*, ya que su valor depende del valor elegido para X.

Puede ocurrir que se asigne el mismo valor U a todos los valores de X, estando formado ahora el conjunto T exclusivamente por un solo elemento. Tenemos entonces un caso especial, en el cual el valor U de la función no varía; es decir, U es *constante*. Incluiremos este caso dentro del concepto general de función, aunque pueda parecer extraño al principiante, a quien, naturalmente, le parece que la importancia del concepto radica en que U varía al variar X. Pero no nos causará ninguna dificultad y, de hecho, será útil el considerar una constante como caso especial de una variable cuyo «dominio de variabilidad» se compone de un solo elemento.

El concepto de función es de capital importancia, no sólo en la matemática pura, sino también en las aplicaciones prácticas. Las leyes de la física no son sino proposiciones respecto a la forma en que dependen ciertas cantidades de otras, cuando algunas de éstas varían. Así, el tono de la nota emitida por una cuerda vibrante depende de su longitud, de su peso y de la tensión a que está sometida; la presión atmosférica depende de la altitud; la energía de una bala, de su masa y de su velocidad. La tarea del físico consiste en determinar la naturaleza exacta o aproximada de esa dependencia funcional.

El concepto de función permite una caracterización exacta del movimiento. Si una partícula en movimiento se considera reducida a

un punto en un espacio de tres dimensiones, siendo x, y, z sus coordenadas rectangulares y designando por t el tiempo, su movimiento queda completamente determinado si se dan sus coordenadas x, y, z como funciones de t:

$$x = f(t), \qquad y = g(t), \qquad z = h(t).$$

Así, si una partícula cae libremente a lo largo del eje vertical z, bajo la sola influencia de la gravedad,

$$x = 0, \qquad y = 0, \qquad z = -\tfrac{1}{2}gt^2,$$

donde g es la aceleración de la gravedad. Si una partícula gira uniformemente en una circunferencia de radio unidad, en el plano x, y, su movimiento está caracterizado por las funciones

$$x = \cos \omega t, \qquad y = \operatorname{sen} \omega t,$$

donde ω es una constante, llamada velocidad angular del movimiento.

Una función matemática no es más que una ley que regula la interdependencia de cantidades variables. No presupone la existencia de una relación de «causa y efecto» entre ellas. Aunque en el lenguaje corriente se utiliza a menudo la palabra «función» con este último sentido, evitaremos todas esas interpretaciones filosóficas; p. ej., la ley de Boyle afirma que para un gas contenido en un recipiente, a temperatura constante, el producto de la presión p y el volumen v es una constante c (cuyo valor depende a su vez de la temperatura); es decir:

$$pv = c.$$

Esta relación permite despejar p en función de v, o v en función de p; $p = c/v$, o bien, $v = c/p$, lo que no significa necesariamente que un cambio de presión sea la «causa» de la variación del volumen, ni tampoco lo contrario. Para el matemático, lo único que importa es la forma de la *relación* entre ambas variables.

Los matemáticos y los físicos difieren, a veces, en cuanto al aspecto del concepto de función, al que conceden gran importancia. Generalmente, los primeros insisten en la *ley de correspondencia*, o sea, la operación matemática que ha de aplicarse a la variable independiente x para obtener el valor de la variable dependiente u. En este sentido, $f(\ )$ es un símbolo que representa una *operación matemática*; el valor $u = f(x)$ es el resultado de aplicar la operación $f(\ )$ al número x. Por otra parte, el físico se interesa generalmente por la *cantidad* u en sí, más que por el procedimiento matemático mediante el cual se obtiene a partir de x. La resistencia u que opone el aire a un cuerpo en movimiento depende de su velocidad v y puede determinarse experimentalmente, sea o no conocida una fórmula matemática explícita para calcular $u = f(v)$. Lo que le interesa primordialmente al físico es la verdadera resistencia y no una fórmula matemática particular $f(v)$,

excepto en cuanto el estudio de dicha fórmula contribuya a analizar el comportamiento de la cantidad u. Es ésta la actitud que se adopta generalmente cuando se *aplica* la matemática a la física o a la ingeniería. En cálculos más avanzados con funciones pueden evitarse muchas veces las confusiones estableciendo exactamente si se considera la operación $f(\ )$, mediante la cual se asigna a x otra cantidad $u = f(x)$, o si interesa primordialmente u, que puede depender de manera enteramente distinta de otra variable z; p. ej., el área de un círculo viene dada por la función $u = f(x) = \pi x^2$, siendo x el radio, y también por la función $u = g(z) = z^2/4\pi$, donde z es la longitud de la circunferencia.

Tal vez, el tipo más sencillo de función matemática de una variable son los *polinomios*, de la forma:

$$u = f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n,$$

con «coeficientes» constantes, $a_0, a_1, a_2, \ldots, a_n$. Después vienen las *funciones racionales*, tales como

$$u = \frac{1}{x}, \qquad u = \frac{1}{1+x^2}, \qquad u = \frac{2x+1}{x^4+3x^2+5},$$

que son cocientes de polinomios; y las *funciones trigonométricas*, sen x, cos x y tg x = sen x/cos x, para las cuales el método más cómodo de definición consiste en referirlas al círculo unidad $\xi^2 + \eta^2 = 1$, en el plano ξ, η. Si el punto $P(\xi, \eta)$ se mueve sobre la circunferencia de este círculo y es x el ángulo que debe girar el semi-eje positivo ξ para que coincida con OP, entonces sen x y cos x son las coordenadas de P: cos $x = \xi$, sen $x = \eta$.

2. Medida de los ángulos en radianes.—En la práctica, los ángulos se miden en unidades que se obtienen dividiendo un ángulo recto en un cierto número de partes iguales. Si son 90, la unidad es el «grado» sexagesimal. Para nuestro sistema decimal sería mejor dividir el ángulo recto en 100 partes iguales, aunque equivaldría al mismo principio de medida. Sin embargo, para ciertos fines teóricos, es ventajoso utilizar un método esencialmente distinto para caracterizar la magnitud de un ángulo y que se llama su «medida en radianes». Numerosas e importantes fórmulas, en las cuales aparecen las funciones trigonométricas de un ángulo, adquieren una forma más sencilla con este sistema que si los ángulos se miden en grados.

Para hallar la medida en radianes de un ángulo, describiremos una circunferencia de radio igual a 1 y centro en el vértice del ángulo. Éste determinará un arco s en la circunferencia, y definimos la longitud de ese arco como la *medida en radianes* del ángulo. Puesto que la circunferencia completa de radio 1 tiene la longitud 2π, el ángulo de 360° tiene como medida 2π radianes. De aquí se deduce que si x

representa la medida en radianes de un ángulo, e y su medida en grados, existe entre ambas la relación $y/360 = x/2\pi$; o sea,

$$\pi y = 180x.$$

Así, un ángulo de 90^o ($y=90$) tiene una medida en radianes $x = 90\pi/180 = \pi/2$, etc. Por otra parte, el ángulo de un radián (ángulo cuya medida en radianes es $x = 1$) es el ángulo que subtiende un arco igual al radio del círculo. Expresado en grados, será igual a $y = 180/\pi = 57{,}2957 \ldots {}^o$. Deberemos multiplicar siempre la medida en radianes x de un ángulo por el factor $180/\pi$ para obtener su medida y en grados.

La medida x de un ángulo en radianes es, por consiguiente, igual al doble del área A del sector del círculo unidad que determina ese ángulo, ya que la razón de dicha superficie a la total del círculo es igual a la del arco correspondiente a la longitud de la circunferencia: $x/2\pi = A/\pi$, $x = 2A$.

En lo que sigue, el ángulo x significará siempre aquel cuya medida en radianes es x. Un ángulo de x grados se escribirá siempre x^o, para evitar ambigüedad.

Es evidente que la medida en radianes es muy útil para las operaciones analíticas. Sin embargo, para los usos prácticos resulta incómoda. Puesto que π es irracional, nunca podríamos volver al mismo punto de la circunferencia, si llevamos repetidas veces el ángulo unidad, es decir, aquel cuya medida en radianes es 1. La unidad usual es tal que, después de llevar 360 veces 1^o, o cuatro veces 90^o, se vuelve al punto de partida.

3. Gráfica de una función. Funciones inversas.—Muchas veces, una simple representación geométrica muestra claramente el carácter de una función. Si x, u, son coordenadas de un plano respecto a un par de ejes perpendiculares, las funciones lineales, tales como $u=ax+b$, están representadas por líneas rectas; las funciones cuadráticas, tales como $u = ax^2 + bx + c$, por parábolas; la función $u = 1/x$, por una hipérbola, etc. Por definición, la *gráfica* de una función cualquiera $u = f(x)$ se compone de todos los puntos del plano cuyas coordenadas x, u cumplen la condición $u = f(x)$. Las funciones sen x, cos x, tg x, están representadas por las curvas de las figuras 151 y 152. Esas gráficas muestran claramente cómo aumentan o disminuyen los valores de la función al variar x.

Un método importante para introducir nuevas funciones es el siguiente: Comenzando por una función conocida $F(X)$, intentemos resolver la ecuación $U = F(X)$, respecto a X, de tal modo que X

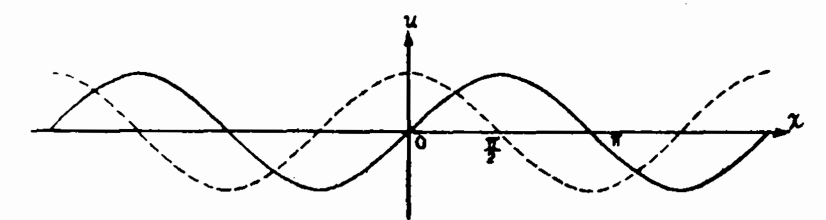


FIG. 151.—Gráficas de sen x y cos x.

aparezca como función de U:

$$X = G(U)$$

La función $G(U)$ se llamará entonces una *función inversa* de $F(X)$. Este procedimiento conduce a un resultado único sólo cuando la función $U = F(X)$ define una representación biunívoca del dominio X en el U; es decir, si la desigualdad $X_1 \neq X_2$ entraña siempre esta otra: $F(X_1) \neq F(X_2)$; pues sólo entonces quedará definida unívocamente una X para cada U.

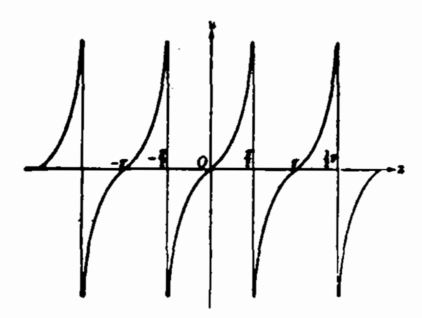


FIG. 152.—$u = \operatorname{tg} x$.

El caso antes visto, en el cual X significaba un triángulo cualquiera del plano, y $U = F(X)$ su perímetro, es un ejemplo oportuno. Es evidente que esta representación del conjunto de los triángulos S en el conjunto T de los números reales positivos no es biunívoca, ya que existen infinitos triángulos que tienen el mismo perímetro. Así, pues, en este caso la relación $U = F(X)$ no sirve para definir una función inversa única. Por otra parte, la función $m = 2n$, donde n varía en todo el conjunto S de los números enteros y m en el conjunto T de los enteros pares, proporciona una correspondencia biunívoca entre ambos conjuntos, y la función inversa $n = m/2$ está definida unívocamente. Otro ejemplo de una representación biunívoca lo proporciona la función

$$u = x^3.$$

Al variar x en el conjunto de todos los números reales, u recorrerá el mismo dominio, tomando cada valor una vez y sólo una. La función inversa, definida unívocamente, es

$$x = \sqrt[3]{u}.$$

En el caso de la función

$$u = x^2,$$

la función inversa no está determinada unívocamente. Puesto que $u = x^2 = (-x)^2$, a cada valor positivo de u corresponderán dos de x. Pero si, como es costumbre, se conviene que el símbolo $\sqrt{u}$ represente el número *positivo* cuyo cuadrado es u, la función inversa

$$x = \sqrt{u}$$

existe, siempre que limitemos la variación de x y de u a valores positivos.

Observando la gráfica de una función $u = f(x)$ de una variable, puede deducirse de una ojeada la existencia de una función inversa única. Ésta estará unívocamente definida, si a cada valor de u corresponde otro de x y uno solo. En la gráfica, esto significa que ninguna paralela al eje x corta a la curva en más de un punto, lo que ocurrirá ciertamente si la función $u = f(x)$ es *monótona;* es decir, si aumenta o disminuye constantemente al aumentar x; p. ej., si $u = f(x)$ es monótona creciente, para $x_1 < x_2$ se tendrá siempre $u_1 = f(x_1) < u_2 = f(x_2)$. De aquí se deduce que para un cierto valor de u puede existir a lo más un x, tal que $u = f(x)$, con lo que la función inversa quedará definida unívocamente. La gráfica de la función inversa $x = g(u)$ se obtiene simplemente haciendo girar el dibujo primitivo un ángulo de 180º alrededor de la bisectriz del primer cuadrante (Fig. 154), con lo cual se cambian entre sí las posiciones de los ejes x y u. La nueva posición de la curva muestra x como función de u. En la posición original de la gráfica, u aparece como la altura por encima del eje horizontal x, mientras que, después de la rotación, la misma curva muestra x como altura sobre el eje horizontal u.

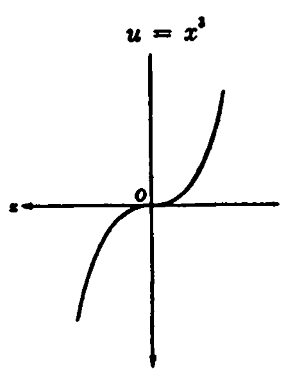


FIG. 153.—$u = x^3$.

Las observaciones del párrafo anterior pueden aclararse considerando el caso de la función

$$u = \operatorname{tg} x.$$

Esta función es monótona para $-\pi/2 < x < \pi/2$ (Fig. 152). Los valores de u, que crecen monótonamente con x, varían desde $-\infty$ a $+\infty$, por lo que la función inversa

$$x = g(u).$$

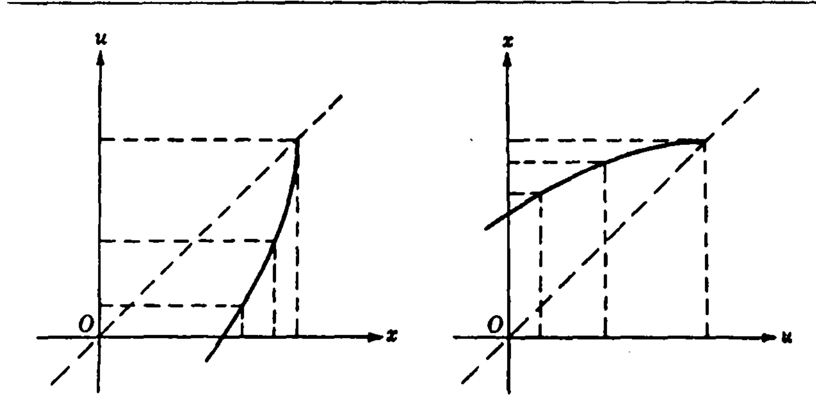


FIG. 154.—Funciones inversas.

está definida para todo valor de u. Se designa por $\operatorname{tg}^{-1}u$, o bien por arc tg u. Así, arc tg $1 = \pi/4$, ya que tg $\pi/4 = 1$. La figura 155 muestra esta función.

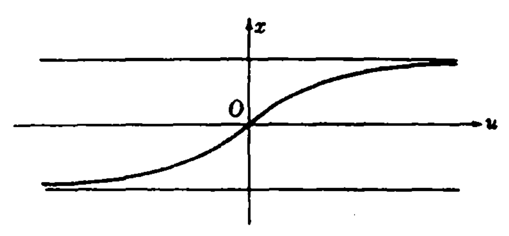


FIG. 155.—$x = \operatorname{arc\,tg} u$.

4. **Funciones compuestas.**—Otro método para crear nuevas funciones, partiendo de dos o más previamente dadas, consiste en formar funciones *compuestas*; p. ej., la función

$$u = f(x) = \sqrt{1 + x^2}$$

se «compone» de estas dos más sencillas:

$$z = g(x) = 1 + x^2, \qquad u = h(z) = \sqrt{z},$$

y puede escribirse de la manera siguiente:

$$u = f(x) = h(g[x]) \qquad \text{(léase «}h\text{ de }g\text{ de }x\text{»).}$$

Análogamente, la función

$$u = f(x) = 1/\sqrt{1 - x^2}$$

está compuesta por las tres funciones:

$$z = g(x) = 1 - x^2, \qquad w = h(z) = \sqrt{z}, \qquad u = k(w) = \frac{1}{w},$$

por lo que resulta:

$$u = f(x) = k(h[g(x)]).$$

La función

$$u = f(x) = \operatorname{sen} \frac{1}{x}$$

se compone de las dos funciones siguientes:

$$z = g(x) = \frac{1}{x} \quad \text{y} \quad u = h(z) = \operatorname{sen} z.$$

La función $f(x)$ no está definida para $x = 0$, puesto que para este valor la expresión $1/x$ no tiene significado. La gráfica de esta función notable se obtiene de la del seno. Sabemos que $\operatorname{sen} z = 0$ para $z = k\pi$, donde k es cualquier entero positivo o negativo. Además,

$$\operatorname{sen} z = \begin{cases} 1 & \text{para} \quad z = (4k+1)\dfrac{\pi}{2}, \\ -1 & \text{para} \quad z = (4k-1)\dfrac{\pi}{2}, \end{cases}$$

siendo k un entero cualquiera. De donde se deduce:

$$\operatorname{sen} \frac{1}{x} = \begin{cases} 0 & \text{para} \quad x = \dfrac{1}{k\pi}, \\ 1 & \text{para} \quad x = \dfrac{2}{(4k+1)\pi}, \\ -1 & \text{para} \quad x = \dfrac{2}{(4k-1)\pi} \end{cases}$$

Si damos sucesivamente a k los valores $k = 1, 2, 3, 4, \cdots$, como los denominadores de estas fracciones aumentan indefinidamente, los valores de x para los que la función $\operatorname{sen}(1/x)$ es igual a $1, -1, 0$, se acumularán cada vez más en torno al punto $x = 0$. Entre uno de esos puntos y el origen existirá además un número infinito de oscilaciones de la función, cuya gráfica aparece en la figura 156.

5. Continuidad.—Las gráficas de las funciones que hemos considerado hasta ahora proporcionan una idea intuitiva del concepto de continuidad. En una sección posterior analizaremos con toda precisión este concepto, después de haber establecido de forma rigurosa el de límite. Hablando en términos generales, diremos que una fun-

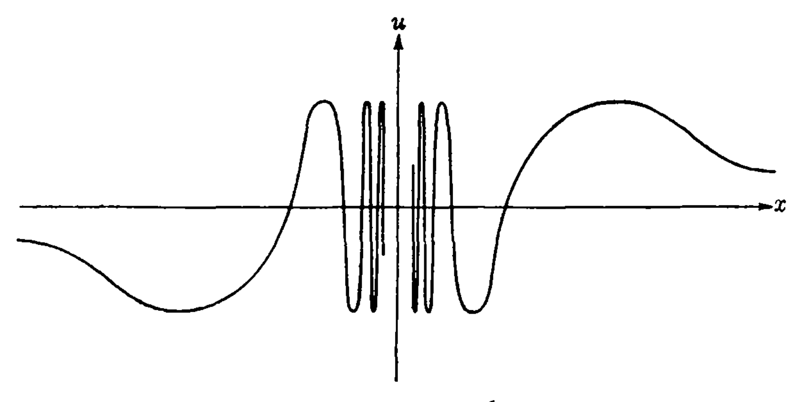


FIG. 156.—$u = \text{sen}\,\frac{1}{x}$

ción es continua cuando la gráfica es una curva sin interrupción (véase pág. 321). Puede estudiarse la continuidad de una función dada, $u = f(x)$, haciendo que la variable independiente x se aproxime indefinidamente, por la derecha y por la izquierda, hacia un valor prefijado x_1. A menos que la función $u = f(x)$ sea constante en el entorno de x_1, su valor cambiará. Si el valor de $f(x)$ tiende al de $f(x_1)$, es decir, al valor de la función en el punto prefijado $x = x_1$, *cualquiera que sea la forma como tienda* x *a* x_1, *por un lado o por el otro,* se dice que la función es *continua* en el punto x_1. Si esto es cierto para todo punto x de un cierto intervalo, se dice que la función es *continua en el intervalo.*

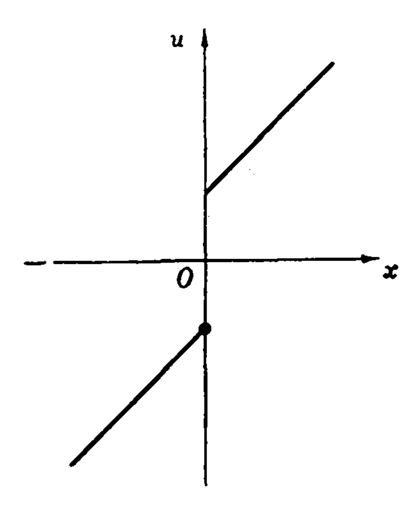
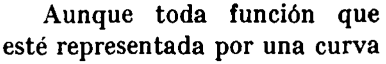


FIG. 157.—Discontinuidad finita.



Aunque toda función que esté representada por una curva sin interrupciones es continua, es fácil definir funciones que no son continuas en todo punto; p. ej., la función de la figura 157, definida para todo valor de x, mediante el convenio

$$f(x) = 1 + x \quad \text{para} \quad x > 0$$
$$f(x) = -1 + x \quad \text{para} \quad x < 0$$

es discontinua en el punto $x_1 = 0$, donde toma el valor -1. Si intentamos dibujar una gráfica de la función, tendremos que levantar el lápiz del papel en este punto. Si nos acercamos al valor $x_1 = 0$ desde la derecha, $f(x)$ tiende a $+1$, que difiere del verdadero valor en ese punto. No basta, para establecer la continuidad, el hecho de que al tender x a 0, por la *izquierda*, $f(x)$ tienda a -1.

La función $f(x)$ definida para todo x por el convenio

$$f(x) = 0 \quad \text{para} \quad x \neq 0, \qquad f(0) = 1,$$

presenta una discontinuidad de distinto tipo en el punto $x_1 = 0$. Aquí, al tender x a 0, existen ambos límites, tanto por la izquierda como por la derecha, y son iguales; pero este límite común no coincide con $f(0)$.

Otro tipo de discontinuidad es el que ofrece la función de la figura 158

$$u = f(x) = \frac{1}{x^2},$$

en el punto $x = 0$. Si se hace tender x a cero, sea por la derecha o por la izquierda, u tiende a infinito; la gráfica de la función se interrumpe en este punto, y pequeñas variaciones de x en el entorno de $x = 0$ pueden producir cambios muy grandes en u. En sentido estricto, el valor de la función no está definido para $x = 0$, ya que infinito no es un número y, en consecuencia, no podemos decir que $f(x)$ sea infinita para $x = 0$. Por tanto, nos limitamos a decir que $f(x)$ «tiende a infinito» al tender x a cero.

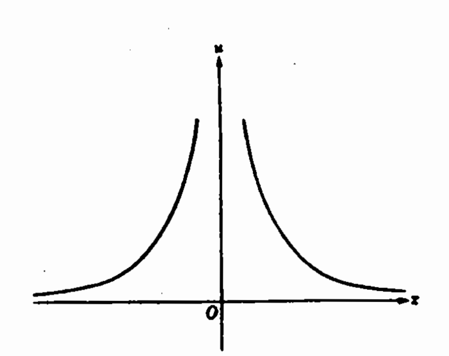


FIG. 158.—Discontinuidad infinita.

Un tipo distinto de discontinuidad aparece en la función $u = \operatorname{sen}(1/x)$, en el punto $x = 0$, como resulta evidente de la gráfica de la función (Fig. 156).

Los ejemplos precedentes ponen de manifiesto las diversas formas en que puede dejar de ser continua una función en un punto $x = x_1$:

1) Es posible a veces conseguir que la función sea continua en $x = x_1$, mediante una definición adecuada de su valor para $x = x_1$; p. ej., la función $u = x/x$ es constantemente igual a 1 si $x \neq 0$; no está, en cambio, definida para $x = 0$, porque $0/0$ es un símbolo ca-

rente de significado. Pero si convenimos en este caso que el valor $u = 1$ corresponde también al valor $x = 0$, la función así generalizada resulta continua para todo valor de x, sin excepción. El mismo efecto se consigue si definimos $f(0) = 0$ para la función que hemos considerado anteriormente. Una discontinuidad de este tipo se dice que es *evitable*.

2) Pueden existir límites distintos cuando x tiende a x_1, según lo haga por la derecha o por la izquierda, como se ve en la figura 157.

3) Es también posible que no existan estos límites laterales, como ocurre en la figura 156.

4) La función puede tender a infinito al tender x a x_1, como en el caso de la figura 158.

Las discontinuidades de los últimos tres tipos se dice que son *esenciales*; no pueden evitarse por una adecuada o nueva definición de la función en el punto $x = x_1$.

Ejercicios:

1. Represéntense las funciones

$$\frac{x-1}{x^2}, \quad \frac{x^2-1}{x^2+1}, \quad \frac{x}{(x^2-1)(x^2+1)}$$

y determínense sus discontinuidades.

2. Dibújense las funciones x sen $(1/x)$ y x^2 sen $(1/x)$ y verifíquese su continuidad para $x = 0$, si se define en ambos casos $u = 0$ para $x = 0$.

*3. Demuéstrese que la función arc tg $(1/x)$ tiene una discontinuidad del segundo tipo (salto), para $x = 0$.

***6. Funciones de varias variables.**—Volvamos a nuestra discusión sistemática del concepto de función. Si la variable independiente P es un punto del plano, de coordenadas x e y, y si a cada punto P corresponde un número único u, que puede ser, p. ej., la distancia del punto P al origen, escribiremos en general

$$u = f(x, y).$$

Se utiliza también esta notación si, como a menudo sucede, las dos cantidades x e y aparecen desde un principio como variables independientes; p. ej., la presión u de un gas es función del volumen x y de la temperatura y, y el área de un triángulo es una función $u = f(x, y, z)$ de las longitudes x, y, z de sus tres lados.

Así como una gráfica nos da la representación geométrica de una función de una variable, una superficie en un espacio de tres dimensiones, de coordenadas x, y, u, permite la representación geométrica de una función $u = f(x, y)$ de dos variables. Asignamos a cada punto

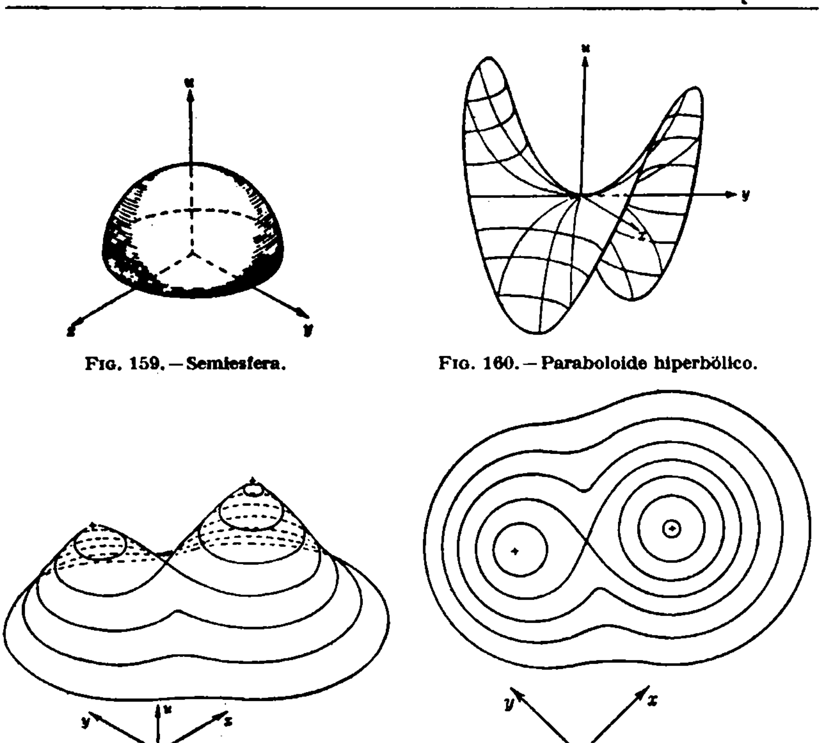


FIG. 159.—Semiesfera. FIG. 160.—Paraboloide hiperbólico.

FIG. 161.—Una superficie $u = f(x, y)$. FIG. 162.—Curvas de nivel correspondientes.

x, y del plano otro punto del espacio, cuyas coordenadas son x, y y $u = f(x, y)$. Así, la función $u = \sqrt{1 - x^2 - y^2}$ está representada por una superficie esférica, de ecuación $u^2 + x^2 + y^2 = 1$; la función lineal $u = ax + by + c$ está representada por un plano, y la función $u = xy$, por un paraboloide hiperbólico, etc.

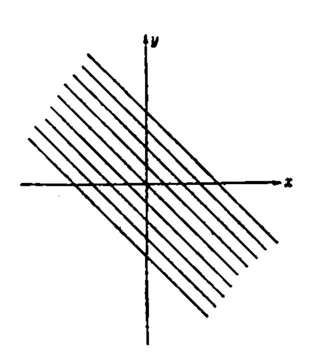


FIG. 163.—Curvas de nivel de $u = x + y$.

Puede darse una representación distinta de la función $u = f(x, y)$ sin salir del plano x, y, mediante *curvas de nivel.* En lugar de considerar el «panorama» tridimensional $u = f(x, y)$, dibujamos, como en un mapa altimétrico, las curvas de nivel de la función, representando las proyecciones de todos los puntos de igual cota u sobre el plano x, y. Estas curvas de nivel son simplemente las curvas

$f(x, y) = c$, en las cuales c permanece constante para cada curva. Así, la función $u = x + y$ queda caracterizada por la figura 163. Las curvas de nivel de una superficie esférica son una familia de circunferencias concéntricas. La función $u = x^2 + y^2$, que representa un paraboloide de revolución, está caracterizada, análogamente, por un sistema de circunferencias (Fig. 165). Por medio de números adscritos a las diferentes curvas puede indicarse su cota, $u = c$.

En física aparecen funciones de varias variables al pretender describir el movimiento de un medio continuo; p. ej., si una cuerda tensa entre dos puntos del eje de las x se deforma de tal modo que la partícula que ocupa el punto x se desplaza perpendicularmente al eje

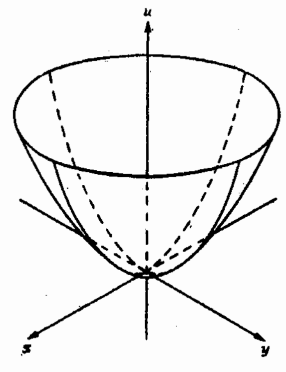


FIG. 164.—Paraboloide de revolución.

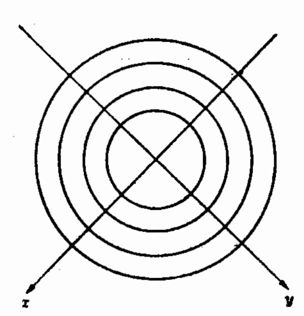


FIG. 165.—Curvas de nivel correspondientes.

una cierta distancia, y se la suelta después, vibrará de tal manera que la partícula cuya abscisa inicial era x, se encontrará en el instante t a una distancia del eje x igual a $u = f(x, t)$. El movimiento queda completamente descrito en cuanto se conoce la función $u = f(x, t)$.

Es posible aplicar a las funciones de varias variables la definición de continuidad dada para las de una sola variable. Se dice que una función $u = f(x, y)$ es continua en el punto $x = x_1$, $y = y_1$, si $f(x, y)$ tiende a $f(x_1, y_1)$ cuando él punto x, y tiende al x_1, y_1, en cualquier forma y desde cualquier dirección.

Existe, sin embargo, una diferencia importante entre las funciones de una y de varias variables. En este último caso, el concepto de función inversa carece de significado, pues es imposible resolver una ecuación $u = f(x, y)$; p. ej., $u = x + y$, de tal manera que cada una de las cantidades independientes x e y aparezca expresada en función de *una sola* cantidad u. Pero esta diferencia en el aspecto de las funcio-

nes de una y de varias variables desaparece si vemos en el concepto de función la idea de una representación o transformación.

***7. Funciones y transformaciones.**—Una correspondencia entre los puntos de una recta l, caracterizados por su abscisa x, y los puntos de otra recta l', determinados por su abscisa x', es simplemente una función $x' = f(x)$. Si esa correspondencia es biunívoca, tendremos también la función inversa $x = g(x')$. El ejemplo más sencillo es el de una transformación por proyección, que, en general, se caracteriza (lo afirmamos sin demostración) por una función de la forma $x' = f(x) = (ax + b)/(cx + d)$, donde a, b, c y d son constantes. En este caso, la función inversa es $x = g(x') = (-dx' + b)/(cx' - a)$.

Las representaciones en dos dimensiones de un plano π, de coordenadas x e y, sobre otro plano π' de coordenadas x' e y', no pueden expresarse por una sola función $x' = f(x)$, sino que se requieren dos funciones de dos variables:

$$x' = f(x, y), \qquad y' = g(x, y).$$

P. ej., una transformación proyectiva está dada por el sistema de funciones

$$x' = \frac{ax + by + c}{gx + hy + k},$$

$$y' = \frac{dx + ey + f}{gx + hy + k},$$

donde $a, b, \ldots, k$, son constantes, y x, y, x', y', son las coordenadas de ambos planos, respectivamente. Desde este punto de vista, la idea de una transformación inversa adquiere perfecto sentido. Simplemente, debemos *resolver este sistema de ecuaciones*, considerando x e y como incógnitas, en función de x' e y'. Geométricamente, esto equivale a encontrar la transformación inversa de π' en π. Estará definida unívocamente si la correspondencia entre los puntos de ambos planos es unívoca.

Las transformaciones del plano estudiadas en topología no están dadas por simples ecuaciones algebraicas, sino por un sistema cualquiera de funciones,

$$x' = f(x, y), \qquad y' = g(x, y),$$

que definen una transformación biunívoca y bicontinua.

Ejercicios:

*1. Demuéstrese que la transformación por inversión (Cap. III, pág. 154) en el círculo unidad está dada analíticamente por las ecuaciones $x' = x/(x^2 + y^2)$, $y' = y/(x^2 + y^2)$. Determínese la transformación inversa. Demuéstrese, analítica-

mente, que la inversión transforma la totalidad de rectas y circunferencias en rectas y circunferencias.

2. Demuéstrese que mediante la transformación $x' = (ax + b)/(cx + d)$ cuatro puntos del eje x se transforman en otros cuatro del eje x', con la misma razón doble (véase pág. 187).

II. LÍMITES

1. Límite de una sucesión a_n.—Como ya hemos visto, el concepto de continuidad de una función se basa sobre el de límite. Hasta ahora hemos utilizado dicho concepto en una forma más o menos intuitiva. En este párrafo y los siguientes lo consideraremos de una manera más sistemática, comenzando por el estudio de las sucesiones, que son más sencillas que las funciones de una variable continua.

En el capítulo II encontramos ya sucesiones de números a_n y estudiamos sus límites, al crecer n indefinidamente o «tender a infinito»; p. ej., la sucesión, cuyo término enésimo es $a_n = 1/n$,

$$1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, \ldots \qquad [1]$$

tiene el límite 0 al crecer n, es decir,

$$\frac{1}{n} \to 0 \quad \text{cuando} \quad n \to \infty \qquad [2]$$

Intentemos formular exactamente lo que se quiere decir con eso; al avanzar en la sucesión, los términos resultan cada vez menores. Después del término de lugar cien, todos son menores que 1/100; después del de lugar mil, menores que 1/1000, etc. Ninguno de ellos es efectivamente igual a cero; pero si *avanzamos suficientemente* en la sucesión, podemos estar seguros de que todos los términos difieren de 0 *tan poco como queramos.*

La única dificultad en esta explicación consiste en que no es completamente claro el significado de las frases en cursiva. ¿Hasta dónde debemos ir para «avanzar suficientemente», y cuánto es «tan poco como queramos»? Si podemos dar un sentido preciso a estas frases, tendrá también un significado preciso la relación [2].

Una interpretación geométrica nos ayudará a aclarar más la cuestión. Si representamos los términos de la sucesión [1] por sus correspondientes puntos sobre una recta, observamos que éstos parecen acumularse alrededor del punto O. Elijamos un intervalo I de centro en el punto O y de amplitud 2ε, de tal modo que se extienda una distancia ε a cada lado del punto O. Si elegimos $\varepsilon = 10$, naturalmente *todos* los términos $a_n = 1/n$ de la sucesión quedarán dentro

de I. Si $\varepsilon = 1/10$, salvo los 10 primeros términos, todos ellos, a partir del a_{11}, es decir,

$$1/11,\ 1/12,\ 1/13,\ 1/14,\ \cdots,$$

quedarán dentro de I. Y si tomamos $\varepsilon = 1/1000$, sólo los primeros mil términos de la sucesión quedarán fuera de I, mientras que a partir del a_{1001}, todos los infinitos términos

$$a_{1001},\ a_{1002},\ a_{1003},\ \cdots$$

serán interiores a I. Es evidente que este razonamiento es válido cualquiera que sea el número positivo ε; una vez elegido éste, por muy pequeño que sea, podemos encontrar un entero N lo bastante grande que

$$1/N < \varepsilon.$$

De aquí se deduce que todos los términos a_n de la sucesión, para los cuales $n \geqslant N$ serán interiores a I y que sólo un número finito de ellos, $a_1, a_2, a_3, \ldots a_{n-1}$, quedará fuera. El punto importante es éste: *primeramente*, se asigna al intervalo I una amplitud arbitraria mediante la elección de ε; *después*, puede encontrarse un entero conveniente N. Este proceso de elegir primero un número ε y encontrar después un entero conveniente N puede efectuarse cualquiera que sea el número positivo ε, por pequeño que se haya tomado, lo que proporciona un significado preciso a la afirmación de que todos los términos de la sucesión [1] difieren de 0 tan poco como se quiera, siempre que se avance suficientemente en la misma.

Resumiendo: sea ε un número positivo cualquiera; entonces, podemos encontrar un entero N, tal que todos los términos a_n de la sucesión [1], para los cuales $n \geqslant N$ son interiores a un intervalo I de amplitud total 2ε y cuyo centro es el punto O. Éste es el significado preciso de la relación [2].

Basándonos en este ejemplo, podemos dar ahora una definición rigurosa de la afirmación general: «la sucesión de números reales a_1, a_2, a_3, ..., tiene el límite a». Incluímos a en el interior de un intervalo I de la recta numérica: si el intervalo es pequeño, algunos de los números a_n quedarán fuera de él; pero apenas sea n suficientemente grande, p. ej., igual o mayor que un cierto entero N, todos los números a_n, para los cuales $n \geqslant N$, serán interiores al intervalo. Naturalmente, es necesario tomar N muy grande, si es muy pequeño el intervalo elegido I; pero cualquiera que sea su amplitud, debe existir un entero N si la sucesión ha de tener como límite a.

Se expresa simbólicamente que una sucesión a_n tiene el límite a de la siguiente manera:

$$\lim a_n = a, \quad \text{cuando} \quad n \to \infty,$$

o simplemente

$$a_n \to a, \quad \text{cuando} \quad n \to \infty$$

(léase: a_n *tiende a* a, o *converge hacia* a).

Puede formularse de forma más concisa la definición de convergencia de una sucesión a_n hacia un límite a, de la manera siguiente: *La sucesión* $a_1, a_2, a_3, \ldots,$ *tiene el límite* a, *cuando* n *tiende a infinito, si para todo número positivo* ε, *por pequeño que sea, existe un entero* N *(que depende de* ε*), tal que*

$$|a - a_n| < \varepsilon \quad \textit{para todo} \quad n \geqslant N. \qquad [3]$$

Ésta es la formulación abstracta del concepto de límite de una sucesión. No es de extrañar que cuando se encuentra por primera vez no sea posible captarla inmediatamente en toda su profundidad. Algunos autores adoptan una actitud poco feliz, presentando esta definición al lector sin una preparación adecuada, como si dar una explicación no resultara muy honroso para la dignidad de un matemático.

La definición sugiere la siguiente disputa entre dos personas A y B. A exige que los a_n se acerquen a la cantidad fijada a con una aproximación mayor que un cierto margen $\varepsilon = \varepsilon_1$. B prueba que satisface esa exigencia demostrando que existe un número entero $N = N_1$ tal que todos los a_n posteriores al elemento a_{N_1} cumplen esa condición. Entonces A se hace más exigente y propone un nuevo margen $\varepsilon = \varepsilon_2$, más pequeño. B accede de nuevo a su demanda encontrando otro entero $N = N_2$ (tal vez mucho mayor). *Si* B *puede satisfacer siempre a* A, *por pequeño que sea el margen que ésta imponga, se produce la situación que hemos expresado mediante* $a_n \to a$.

Existe una dificultad psicológica auténtica para comprender esta definición precisa de límite. Nuestra intuición nos sugiere una idea «dinámica» del límite, como si fuera el resultado de un «movimiento»: recorremos la sucesión de enteros 1, 2, 3, ..., n, ... , y observamos el comportamiento de la sucesión a_n. Sentimos que debe ser posible observar la convergencia $a_n \to a$; pero es imposible dar una formulación matemática clara de esta actitud «natural». Para llegar a una definición precisa, tenemos que *invertir* el orden del proceso; en lugar de observar primero la variable independiente n y después la dependiente a_n, debemos basar nuestra definición en lo que es necesario hacer si queremos verificar la afirmación según la cual $a_n \to a$. Para

proceder de esa manera, debemos elegir primeramente una distancia arbitrariamente pequeña en torno de a, y determinar si podemos satisfacer esa condición eligiendo un valor suficientemente grande de la variable independiente n. Dando después denominaciones simbólicas, ε y N, a las palabras «arbitrariamente pequeño» y «n suficientemente grande», llegamos a la definición precisa de límite.

Consideremos como nuevo ejemplo la sucesión

$$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \ldots, \frac{n}{n+1}, \ldots,$$

en la cual, $a_n = n/(n+1)$. Digo que el límite es 1. Si el lector elige un intervalo cuyo centro sea el punto 1 y para el cual $\varepsilon = 1/10$, puedo satisfacer su exigencia [3] tomando $N = 10$, pues

$$0 < 1 - \frac{n}{n+1} = \frac{n+1-n}{n+1} = \frac{1}{n+1} < \frac{1}{10}$$

en cuanto $n \geqslant 10$. Si el lector se hace más exigente y elige $\varepsilon = 1/1000$, puedo satisfacer esa demanda tomando $N = 1000$, y análogamente, para cualquier número positivo ε, por pequeño que sea, pues me bastará elegir un entero N mayor que $1/\varepsilon$. Este proceso de asignar una distancia arbitrariamente pequeña ε alrededor del número a y demostrar después que los términos de la sucesión a_n están todos dentro de una distancia ε de a, si avanzamos en ella suficientemente, es la descripción minuciosa del hecho de que $\lim a_n = a$.

Si los términos de la sucesión $a_1, a_2, a_3, \ldots$, se expresan en forma decimal, la afirmación $\lim a_n = a$ significa simplemente que para cualquier entero positivo m, las primeras m cifras de a_n coinciden con las m primeras del desarrollo decimal del número fijo a, con tal que se tome n suficientemente grande; p. ej., mayor o igual que un cierto entero N (que depende de m). Esto corresponde simplemente a elegir ε en la forma 10^{-m}.

Existe otro método muy sugestivo de expresar el concepto de límite. Si $\lim a_n = a$, y encerramos a en el interior de un intervalo I, entonces, por pequeño que sea I, todos los números a_n, para los cuales n es igual o mayor que un entero N, se encontrarán dentro de I, por lo que, contando desde el principio de la sucesión, a lo sumo *un número finito*, N — 1, *de términos*

$$a_1, a_2, \cdots, a_{N-1},$$

pueden estar fuera de I. Si I es muy pequeño, N será muy grande, p. ej., cien o mil millones, pero siempre quedará fuera de I sólo un

número finito de términos, mientras que los infinitos restantes serán interiores a I.

Podemos decir que *casi todos* los términos de una cierta sucesión tienen una determinada propiedad si sólo un número finito de ellos, no importa lo grande que sea, no la poseen; p. ej., «casi todos» los enteros positivos son mayores que 1 000 000 000 000. Utilizando esta terminología, la afirmación $\lim a_n = a$ equivale a decir que *si* I *es un intervalo cualquiera cuyo centro es* a, *casi todos los números* a_n *se encuentran dentro de* I.

Conviene observar de pasada que no debe suponerse que todos los términos a_n de la sucesión han de tener necesariamente valores distintos. Es posible que algunos, infinitos, o incluso *todos* ellos, sean iguales al valor límite a; p. ej., la sucesión $a_1 = 0$, $a_2 = 0$, ..., $a_n = 0$, ... es perfectamente legítima y, naturalmente, su límite es cero.

Una sucesión a_n que tiene un límite a se llama *convergente*. Una sucesión a_n que carece de límite se llama *divergente*.

Ejercicios: Demuéstrese:

1. La sucesión $a_n = n/(n^2 + 1)$ tiene límite 0. (*Indicación*: $a_n = 1/(n + 1/n)$ es menor que $1/n$ y mayor que 0.)

2. La sucesión $a_n = (n^2 + 1)/(n^3 + 1)$ tiene el límite 0. (*Indicación*: $a_n = (1 + 1/n^2)/(n + 1/n^2)$ está comprendido entre 0 y $2/n$.)

3. La sucesión 1, 2, 3, 4, $\cdots$ y las sucesiones oscilantes

$$1, 2, 1, 2, \cdots,$$
$$1, -1, 1, -1, \cdots \text{ es decir, } a_n = (-1)^n,$$
$$1, 1/2, 1, 1/3, 1, 1/4, 1, 1/5, \cdots,$$

no tienen límite.

Si en una sucesión a_n los términos llegan a ser mayores que cualquier número prefijado K, diremos que a_n *tiende a infinito* y escribiremos: $\lim a_n = \infty$, o $a_n \to \infty$; p. ej., $n^2 \to \infty$ y $2^n \to \infty$. Esta terminología es útil, aunque quizá no del todo satisfactoria, puesto que no puede considerarse ∞ como un número a. *Una sucesión que tiende a infinito se llama también divergente.*

Ejercicios: Demuéstrese que las sucesiones $a_n = \dfrac{n^2+1}{n}$, $a_n = \dfrac{n^2+1}{n+1}$, $a_n = \dfrac{n^3-1}{n+1}$ y $a_n = \dfrac{n^n}{n^2+1}$ tienden todas a infinito.

A veces, los principiantes cometen el error de creer que un paso al límite, cuando $n \to \infty$, se lleva a cabo simplemente sustituyendo n por ∞ en la expresión de a_n; p. ej., $1/n \to 0$, puesto que «$1/\infty = 0$». Pero el símbolo ∞ no es un número, y su utilización en la expresión

$1/\infty$ es ilegítima. Imaginarse el límite de una sucesión como si fuera el «último» término a_n, cuando $n = \infty$, significa desconocer el verdadero sentido del concepto y oscurece sus consecuencias.

2. Sucesiones monótonas.—En la definición general dada en la página 303 no se exigió una forma determinada para que la sucesión $a_1, a_2, a_3, \ldots$, tendiera a su límite. El ejemplo más sencillo es el de las sucesiones que se llaman monótonas, tales como

$$1/2,\ 2/3,\ 3/4,\ \cdots,\ n/(n+1)\ \cdots.$$

Cada término de esta sucesión es mayor que el precedente, pues $a_{n+1} = \frac{n+1}{n+2} = 1 - \frac{1}{n+2} > 1 - \frac{1}{n+1} = \frac{n}{n+1} = a_n$. Se denominan *monótonas crecientes* las sucesiones de este tipo para las cuales $a_{n+1} > a_n$. Análogamente, una sucesión para la cual $a_n > a_{n+1}$ como, p. ej., 1, 1/2, 1/3, ..., se llama *monótona decreciente*. Las sucesiones de este tipo se aproximan a su límite exclusivamente desde un lado; en contraste con ello, existen sucesiones que oscilan tal como: -1, $+1/2$, $-1/3$, $+1/4$, ..., que se aproxima a su límite 0 desde ambos lados (véase Fig. 11).

Es fácil determinar el comportamiento de una sucesión monótona; puede carecer de límite y exceder a cualquier número, como, p. ej.,

$$1,\ 2,\ 3,\ 4,\ \cdots,$$

en la cual $a_n = n$, o la sucesión

$$2,\ 3,\ 5,\ 7,\ 11,\ 13,\ \cdots,$$

donde a_n es el n-ésimo número primo, p_n.

En este caso la sucesión tiende a infinito; pero si los términos de una sucesión monótona creciente están acotados; es decir, si todo término es menor que una cota superior B, conocida de antemano, resulta intuitivo que debe tender a un cierto límite a, que será menor,

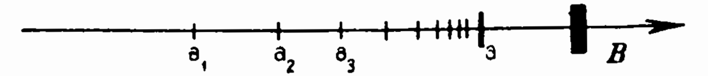


FIG. 166.—Sucesión monótona acotada.

o a lo sumo igual a B. Esta proposición, que llamaremos *principio de las sucesiones monótonas*, se enuncia así: *Toda sucesión monótona creciente, que está acotada superiormente, tiene límite.* (Puede hacerse una afirmación análoga para una sucesión monótona *decreciente*, que esté acotada *inferiormente.*) Es notable que no haga falta conocer

previamente el valor del límite a; el teorema afirma que, si se cumplen las condiciones impuestas, tal límite existe. Naturalmente, este teorema depende de la introducción de los números irracionales, pues de otra manera no sería siempre cierto. Como ya hemos visto en el capítulo II, cualquier número irracional (p. ej., $\sqrt{2}$) es límite de una sucesión monótona creciente y acotada de fracciones racionales decimales, cuyos términos se obtienen tomando las n primeras cifras de una fracción decimal de infinitas cifras.

*Aunque el principio de las sucesiones monótonas resulta evidente a la intuición, es instructivo dar una demostración rigurosa del mismo. Vamos a ver que dicho principio es una consecuencia lógica de las definiciones de número real y de límite.

Supongamos que los números a_1, a_2, a_3, ... forman una sucesión creciente, pero acotada. Podemos expresarlos en forma de fracciones decimales de infinitas cifras.

$$\begin{aligned} a_1 &= A_1,\, p_1p_2p_3 \cdots, \\ a_2 &= A_2,\, q_1q_2q_3 \cdots, \\ a_3 &= A_3,\, r_1r_2r_3 \cdots, \\ &\cdots\cdots\cdots\cdots \end{aligned}$$

donde los A_i representan números enteros y los p_i, q_i, etc., son cifras de 0 a 9. Puesto que la sucesión a_1, a_2, a_3... está *acotada*, estos enteros no pueden crecer indefinidamente, y como la sucesión es *monótona creciente* la sucesión de los enteros A_1, A_2, $A_3 \cdots$, *permanecerá constante después de alcanzar su valor máximo.* Sea A dicho máximo y supongamos que lo alcance en la N_0-ésima fila. Recorramos ahora la segunda columna p_1, q_1, r_1, $\cdots$, limitando nuestra atención a los términos de la N_0-ésima fila y las siguientes. Si x_1 es la mayor cifra que aparece en esta columna después de la N_0-ésima fila, x aparecerá *reiteradamente* después de su primera aparición, lo que puede ocurrir en la N_1-ésima fila, siendo $N_1 > N_0$, pues si las cifras de esta columna disminuyeran después de esta fila, la sucesión a_1, $a_2, a_3, \cdots$, no sería monótona creciente. Consideremos ahora las cifras p_2, q_2, $r_2, \cdots$, de la tercera columna. Un razonamiento análogo demuestra que después de un cierto entero $N_2 \geqslant N_1$, las cifras de la tercera columna son constantemente iguales a x_2. Repitiendo este proceso para las 4.ª, 5.ª, $\cdots$, columnas, obtenemos los dígitos x_3, x_4, $x_5 \cdots$ y una sucesión correspondiente de números enteros N_3, N_4, N_5, $\cdots$ Es fácil ver que el número

$$a = A,\, x_1x_2x_3x_4 \cdots$$

es el límite de la sucesión a_1, a_2, a_3... Pues si se elige un $\varepsilon \geqslant 10^{-m}$, para todos los $n \geqslant N_m$, la parte entera y las m primeras cifras después de la coma de a_m coincidirán con las de a, por lo que la diferencia $|a - a_n|$ no puede exceder a 10^{-m}. Puesto que es posible lograr esto para cualquier ε positivo, por pequeño que sea, ya que bastará tomar m suficientemente grande, el teorema queda demostrado.

Se llega a este mismo resultado basándonos en cualquiera de las otras definiciones de número real, dadas en el capítulo II; p. ej., mediante los encajes de intervalos o las cortaduras de Dedekind. Estas demostraciones pueden encontrarse en cualquier tratado de análisis matemático.

Este principio de las sucesiones monótonas podía haberse utilizado en el capítulo II para definir la suma y el producto de dos números con infinitas cifras decimales:

$$a = A, a_1a_2a_3 \cdots,$$
$$b = B, b_1b_2b_3 \cdots$$

Es imposible sumar o multiplicar dos expresiones de ese tipo de la manera corriente, es decir, empezando por la última cifra decimal de la derecha, pues no existe tal última cifra. (Como ejemplo, el lector puede intentar sumar los dos números decimales 0,333333 ··· y 0,989898 ···). Pero si x_n representa la fracción decimal *finita*, obtenida tomando n cifras decimales en a y b, y sumando de la manera ordinaria, la sucesión $x_1, x_2, x_3, \cdots$ será monótona creciente y estará acotada (p. ej., por el entero $A + B + 2$). De ahí que esta sucesión tenga límite y podemos definir $a + b = \text{lím } x_n$. Un método análogo sirve para definir el producto ab. Mediante las leyes de cálculo de la aritmética es posible extender estas definiciones hasta abarcar todos los casos, cualesquiera que sean los signos de a y b.

Ejercicio: Demuéstrese de esta manera que la suma de los dos números considerados antes es 1,323232 ··· = 131/99.

La importancia del concepto de límite en matemática reside en que *muchos números se definen exclusivamente mediante límites*, a menudo como límites de sucesiones monótonas acotadas. Por esta razón, el campo de los números racionales, en el cual dichos límites pueden no existir, resulta demasiado restringido para las necesidades de la matemática.

3. El número «e» de Euler.—El número e ocupa un lugar destacado en la matemática, junto con el número π de Arquímedes, desde la publicación, en 1748, de la obra de Euler *Introductio in Analysin Infinitorum*. Proporciona un excelente ejemplo del modo en que el principio de las sucesiones monótonas puede servir para definir un nuevo número real. Utilizando el símbolo

$$n! = 1 \cdot 2 \cdot 3 \cdot 4 \cdots n$$

para el producto de los n primeros números enteros, estudiemos la sucesión a_1, a_2, a_3, en la cual

$$a_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!} \qquad [4]$$

Los términos a_n forman una sucesión monótona creciente, puesto que se obtiene a_{n+1} sumando a a_n una cantidad positiva $1/(n + 1)!$. Además, los números a_n están acotados superiormente:

$$a_n < B = 3, \qquad [5]$$

pues se tiene:

$$\frac{1}{s!} = \frac{1}{2}\frac{1}{3}\cdots\frac{1}{s} < \frac{1}{2}\frac{1}{2}\cdots\frac{1}{2} = \frac{1}{2^{s-1}},$$

de donde resulta:

$$a_n < 1 + 1 + \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots + \frac{1}{2^{n-1}} = 1 + \frac{1-(1/2)^n}{1-1/2} =$$

$$= 1 + 2(1-(1/2)^n) < 3,$$

utilizando la fórmula dada en la página 20 para la suma de los n primeros términos de una progresión geométrica. De acuerdo con el principio de las sucesiones monótonas, cuando n tiende a ∞, a_n debe tender a un límite, y *a este límite lo llamaremos* e. Para expresar que $e = \text{lím}\, a_n$, podemos escribir e en forma de serie:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!} + \cdots. \qquad [6]$$

Esta «igualdad», que incluye una sucesión de puntos suspensivos, es simplemente otro método de expresar las dos afirmaciones siguientes:

$$a_n = 1 + \frac{1}{1!} + \frac{1}{2!} + \cdots + \frac{1}{n!}$$

y

$$a_n \to e \quad \text{cuando} \quad n \to \infty.$$

La serie [6] permite calcular e con la aproximación que se desee; p. ej., la suma (con ocho cifras decimales) de los términos de [6] hasta $1/12!$ inclusive es $\Sigma = 2{,}71828183...$ (El lector debe verificar este resultado.) Es fácil estimar el «error»; esto es, la diferencia entre ese valor y el verdadero de e. Para la diferencia $(e - \Sigma)$ tenemos la expresión

$$\frac{1}{13!} + \frac{1}{14!} + \cdots < \frac{1}{13!}\left(1 + \frac{1}{13} + \frac{1}{13^2} + \cdots\right) =$$

$$= \frac{1}{13!}\cdot\frac{1}{1-\frac{1}{13}} = \frac{1}{12\cdot 12!}$$

Esta diferencia es tan pequeña que no puede afectar a la octava cifra decimal de Σ. Suponiendo que exista un posible error en la última cifra del valor dado más arriba, tenemos $e = 2{,}7182818$, con siete cifras decimales exactas.

******El número* e *es irracional.*—Para demostrarlo procederemos indirectamente, suponiendo que fuera $e = p/q$, donde p y q son enteros, y viendo que esta hipótesis conduce a un absurdo. Puesto que sabemos que $2 < e < 3$, e no puede ser entero, por lo que q debe ser, por lo menos, igual a 2. Multiplicando ambos miembros de [6] por $q! = 2 \cdot 3 \cdots q$, se tiene

$$e \cdot q! = p \cdot 2 \cdot 3 \cdots (q-1) = [q! + q! + 3 \cdot 4 \cdots q + 4 \cdot 5 \cdots q + \\ + \cdots + (q-1)q + q + 1] + \frac{1}{(q+1)} + \frac{1}{(q+1)(q+2)} + \cdots . \quad [7]$$

Evidentemente, el primer miembro es un número entero; en el segundo miembro, la expresión entre corchetes lo es igualmente; pero los restantes términos tienen suma positiva menor que $^1/_2$, por lo que no pueden componer un entero, ya que $q > 2$ y los términos de la serie $1/(q+1) + \cdots$ son menores que los correspondientes de la serie geométrica $1/3 + 1/3^2 + 1/3^3 \cdots$, cuya suma es $1/2$. De aquí se deduce que [7] entraña una contradicción: el entero del primer miembro no puede ser igual al valor del segundo, pues éste no es un entero, por cuanto se compone de un entero y de un número positivo menor que $1/2$.

4. El número π.—Como se sabe desde el bachillerato, la longitud de la circunferencia de radio 1 puede definirse como el límite de la sucesión de los perímetros de polígonos regulares de un número creciente de lados. La longitud de la circunferencia así definida se representa mediante el símbolo 2π. Con más precisión, si p_n indica el perímetro del polígono inscrito y q_n el del polígono circunscrito, se tiene $p_n < 2\pi < q_n$. Además, al crecer n, cada una de esas sucesiones p_n y q_n tiende monótonamente a 2π, por lo que, con cada paso, disminuye el error de la aproximación a 2π que representan p_n o q_n.

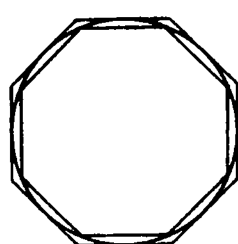


Fig. 167.—Aproximaciones de la circunferencia mediante polígonos.

En la página 135 encontramos la expresión

$$p_{2^m} = 2^m \sqrt{2 - \sqrt{2 + \sqrt{2 + \cdots}}}$$

que contiene $m - 1$ signos de raíces cuadradas, fórmula que puede utilizarse para calcular un valor aproximado de 2π.

Ejercicios:

1. Determínese el valor aproximado de π que proporcionan p_4, p_8 y p_{16}.

*2. Hállese una fórmula para q_{2^m}.

*Utilícese esa fórmula para calcular q_4, q_8 y q_{16}. Habiendo calculado p_{16} y q_{16}, determínese una cota superior y otra inferior de π.

¿Qué *es* el número π? La desigualdad $p_n < 2\pi < q_n$ nos da una respuesta completa, pues determina un encaje de intervalos que definen el punto 2π. Sin embargo, esa respuesta deja algo que desear, pues no nos dice nada acerca de la naturaleza de π como número real; ¿es racional, irracional, algebraico o trascendente? Ya hemos dicho en la página 152 que π es un número trascendente y, por consiguiente, irracional. Al revés de lo que ocurre con e, la demostración de la irracionalidad de π, que J. H. Lambert (1728-1777) dió por primera vez, es bastante difícil, por lo que no la intentaremos reproducir aquí. Sin embargo, hay algunas otras propiedades referentes a π que están a nuestro alcance. Recordando aquella afirmación de que los números enteros son el material básico de la matemática, podemos preguntarnos si el número π tiene alguna relación sencilla con los enteros. Aunque se ha calculado la expresión decimal de π con varios centenares de cifras, no aparece en ellas ninguna regularidad, lo que no es sorprendente, puesto que π y 10 nada tienen de común. Pero en el siglo XVIII, Euler y otros matemáticos encontraron bellísimas expresiones, mediante series y productos infinitos, que establecen un lazo de unión entre los números enteros y π. Tal vez la más sencilla de todas estas fórmulas sea la siguiente:

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots,$$

que expresa $\pi/4$ como límite de la sucesión de sumas parciales:

$$s_n = 1 - \frac{1}{3} + \frac{1}{5} - \cdots + (-1)^n \frac{1}{2n+1}$$

Demostraremos esta fórmula en el capítulo VIII. Otra serie relativa al número π es la siguiente:

$$\frac{\pi^2}{6} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \frac{1}{5^2} + \frac{1}{6^2} + \cdots;$$

Una notable expresión de π fué descubierta por el matemático inglés John Wallis (1616-1703); su fórmula dice que

$$\left\{\frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdots\frac{2n}{2n-1}\cdot\frac{2n}{2n+1}\right\} \to \frac{\pi}{2} \text{ cuando } n \to \infty,$$

lo que a veces se escribe abreviadamente así:

$$\frac{\pi}{2} = \frac{2}{1}\cdot\frac{2}{3}\cdot\frac{4}{3}\cdot\frac{4}{5}\cdot\frac{6}{5}\cdot\frac{6}{7}\cdot\frac{8}{7}\cdot\frac{8}{9}\cdots$$

La expresión que figura en el segundo miembro se llama un *producto infinito.*

La demostración de estas dos fórmulas puede consultarse en cualquier tratado de análisis matemático.

***5. Fracciones continuas.**—Las fracciones continuas conducen a casos interesantes de pasos al límite. Una fracción continua, tal como

$$\frac{57}{17} = 3 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{5}}},$$

representa un número racional. Vimos en la página 57 que todo número racional puede escribirse en esa forma por medio del algoritmo de Euclides. Sin embargo, para los números irracionales, el algoritmo no termina después de un número finito de pasos, sino que conduce a una sucesión de fracciones de longitud creciente, cada una de las cuales representa un número racional. En particular, todos los números reales algebraicos (véase pág. 112) de grado 2 pueden expresarse de este modo. Como ejemplo, sea el número $x = \sqrt{2} - 1$, raíz de la ecuación cuadrática

$$x^2 + 2x = 1, \qquad x = \frac{1}{2 + x}$$

Si en el segundo miembro se sustituye x por $1/(2 + x)$, se obtiene la expresión

$$x = \cfrac{1}{2 + \cfrac{1}{2 + x}},$$

y a continuación,

$$x = \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + x}}},$$

etcétera.

De forma que después de reiterar n veces obtenemos la ecuación

$$x = \left.\cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \ddots \; \cfrac{1}{2 + x}}}}\right\} (n \text{ veces})$$

Al hacer tender n a infinito se obtiene la «fracción continua indefinida»

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \ddots}}}}$$

Esta notable fórmula relaciona $\sqrt{2}$ con los números enteros de forma más sorprendente que lo hace la expresión decimal de $\sqrt{2}$, en la cual no aparece ninguna regularidad en la sucesión de sus cifras.

Para la raíz positiva de cualquier ecuación cuadrática de la forma

$$x^2 = ax + 1, \qquad \text{o} \qquad x = a + \frac{1}{x},$$

se tiene el desarrollo

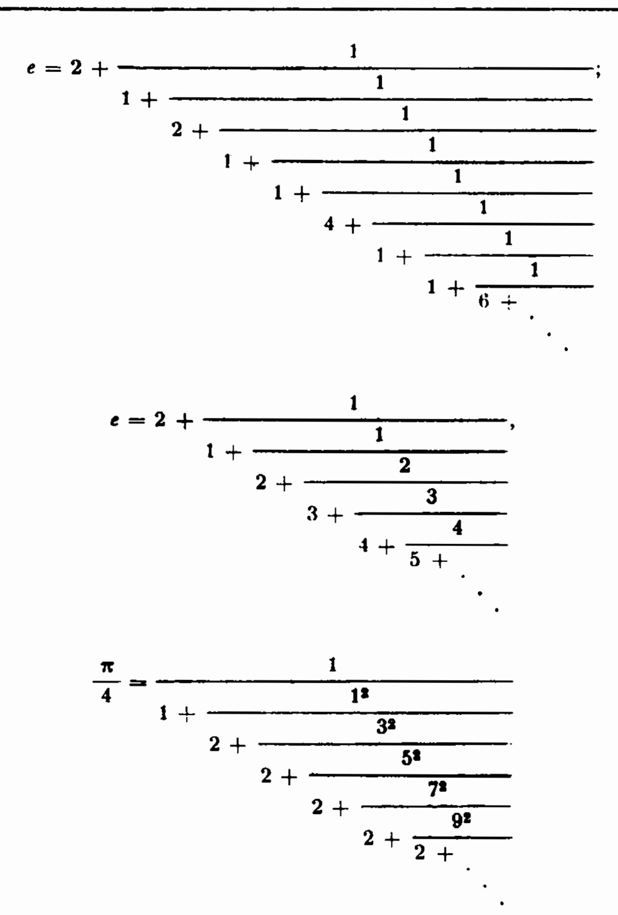
$$x = a + \cfrac{1}{a + \cfrac{1}{a + \cfrac{1}{a + \ddots}}}$$

P. ej., para $a = 1$, obtenemos:

$$x = {}^1/_2(1 + \sqrt{5}) = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{1 + \ddots}}}$$

Estos ejemplos son casos particulares de un teorema general, según el cual *las raíces reales de una ecuación cuadrática de coeficientes enteros admiten desarrollos en fracción continua periódica*, exactamente como los números racionales poseen desarrollos decimales periódicos.

Euler encontró desarrollos en fracción continua para e y π, casi tan sencillos como los anteriores. Daremos los siguientes, sin demostración:



$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{4 + \cfrac{1}{1 + \cfrac{1}{1 + \cfrac{1}{6 + \cdots}}}}}}}};$$

$$e = 2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{2}{3 + \cfrac{3}{4 + \cfrac{4}{5 + \cdots}}}}},$$

$$\frac{\pi}{4} = \cfrac{1}{1 + \cfrac{1^2}{2 + \cfrac{3^2}{2 + \cfrac{5^2}{2 + \cfrac{7^2}{2 + \cfrac{9^2}{2 + \cdots}}}}}}$$

III. LÍMITES POR APROXIMACIÓN CONTINUA

1. Introducción. Definición general.—En la página 301 y sucesivas pudimos dar una formulación precisa de la afirmación siguiente: «la sucesión a_n (es decir, la función $a_n = F(n)$ de la variable entera n) tiene el límite a cuando n tiende a infinito». Daremos ahora la definición correspondiente a esta otra afirmación: «la función $u = f(x)$ de la variable continua x tiene el límite a cuando x tiende a x_1». En forma intuitiva, utilizamos ya este concepto de límite por aproximación continua de la variable independiente x en la página 295, para comprobar la continuidad de la función $f(x)$.

Comenzaremos, como allí, por un caso particular. La función $f(x) = (x + x^3)/x$ está definida para todos los valores de x, salvo el $x = 0$, pues en este punto se anula el denominador. Al trazar la gráfica de la función $u = f(x)$ para valores de x del entorno de 0, parece evidente que, al «acercarse» x a cero por ambos lados, el correspondiente valor de $u = f(x)$ «se aproxima» al límite 1. Para dar una descripción precisa de este hecho debemos encontrar una fórmula explícita de la diferencia entre el valor de $f(x)$ y el número fijo 1:

$$f(x) - 1 = \frac{x + x^3}{x} - 1 = \frac{x + x^3 - x}{x} = \frac{x^3}{x}$$

Si convenimos en considerar exclusivamente valores de x próximos a 0, pero no el valor $x = 0$ [para el cual no está definida $f(x)$], podemos dividir numerador y denominador del segundo miembro de esta ecuación por x, obteniendo la fórmula más sencilla

$$f(x) - 1 = x^2.$$

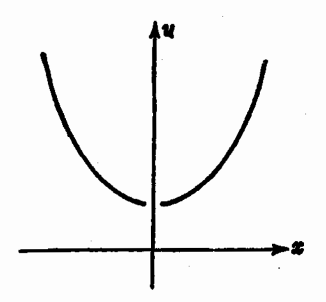


FIG. 168.—$u = (x + x^3)/x$.

Es evidente que esta diferencia puede llegar a ser *tan pequeña como deseemos*, siempre que x esté limitado en un intervalo *suficientemente pequeño* del entorno de 0. Así, para $x = \pm 1/10$, $f(x) - 1 = 1/100$; para $x = \pm 1/100$, $f(x) - 1 = 1/10000$, etc. En general, si ε es un número positivo, por pequeño que sea, la diferencia entre $f(x)$ y 1 será menor que ε, con tal que la distancia de x a 0 sea menor que el número $\delta = \sqrt{\varepsilon}$. Pues entonces, si $|x| < \sqrt{\varepsilon}$, resulta $|f(x) - 1| = |x|^2 < \varepsilon$.

Es completa la analogía con nuestra definición de límite; en la página 303 decíamos que «la sucesión a_n tiene el límite a cuando n tiende a infinito, si en correspondencia con todo número positivo ε, por pequeño que sea, puede encontrarse un entero N (que depende de ε), y tal que $|a_n - a| < \varepsilon$ para todos los n que satisfacen la desigualdad $n \geqslant N$».

En el caso de una función $f(x)$ de una variable continua x, cuando ésta tiende a un valor finito x_1, nos limitamos a reemplazar el n «suficientemente grande», dado por N, por un x_1 «suficientemente próximo», determinado por un número δ, llegando así a la siguiente definición de límite por aproximación continua, que Cauchy fué el primero en enunciar, alrededor de 1820: *La función* f(x) *tiene el límite* a *cuando*

x *tiende a* x_1 *si, para todo número positivo* ε, *por pequeño que sea, puede encontrarse otro* δ *(también positivo y dependiente de* ε*), tal que*

$$| f(x) - a | < \varepsilon$$

para todo $x \neq x_1$ *que satisfaga la desigualdad*

$$| x - x_1 | < \delta.$$

En este caso, escribiremos:

$$f(x) \to a \qquad \text{cuando} \qquad x \to x_1.$$

En el caso de la función $f(x) = (x + x^3)/x$, demostramos anteriormente que $f(x)$ tiende al límite 1 cuando x tiende al valor $x_1 = 0$. En dicho caso bastó elegir $\delta = \sqrt{\varepsilon}$.

2. Observaciones sobre el concepto de límite.—La definición (ε, δ) de límite es el producto de más de un siglo de intentos y encierra en unas pocas palabras el resultado de los persistentes esfuerzos para establecer este concepto sobre una base matemática firme. Los conceptos fundamentales del cálculo—los de derivada e integral—pueden definirse únicamente mediante pasos al límite. Una dificultad, al parecer insuperable, bloqueó durante mucho tiempo el camino hacia la comprensión clara y la definición precisa del concepto de límite.

En el estudio del movimiento, los matemáticos de los siglos XVII y XVIII aceptaron como algo natural el concepto de cantidad x, que cambia continuamente y tiende, en un flujo continuo, hacia un valor límite x_1. Asociado a este flujo primario del tiempo, o de una cantidad x que se comportaba como él, consideraron un valor secundario, $u = f(x)$, que seguía el movimiento de x. El problema consistía en dar un significado matemático preciso a la idea de que $f(x)$ «tiende» o «se aproxima» a un valor fijo a, cuando x tiende hacia x_1.

Desde la época de Zenón y de sus paradojas, el concepto intuitivo, físico o metafísico, de la continuidad del movimiento ha eludido todas las tentativas de una formulación matemática precisa. No hay ninguna dificultad en proceder, paso a paso, a través de una sucesión discreta de valores a_1, a_2, a_3, ... ; pero, al estudiar una variable continua x que se extiende por todo un intervalo de la recta numérica, es imposible decir cómo ha de «aproximarse» x a un valor fijo x_1, de tal manera que tome consecutivamente y en su orden de magnitud todos los valores del intervalo. Pues los puntos de una recta forman un conjunto denso, y no existe un punto «siguiente» después de haber alcanzado uno prefijado. Ciertamente, la idea intuitiva del con-

tinuo posee una realidad psicológica en la mente humana; pero es imposible recurrir a ella para resolver una dificultad matemática; debe subsistir una discrepancia entre la idea intuitiva y el lenguaje matemático ideado para describir los rasgos científicamente importantes de nuestra intuición en una nomenclatura lógicamente precisa. Las paradojas de Zenón son una importante indicación de esta discrepancia.

El mérito de Cauchy consistió en comprender que, en lo que respecta a los conceptos matemáticos, puede y debe omitirse cualquier referencia a una idea intuitiva previa de la continuidad del movimiento. Como ocurre a menudo, se abrió una senda hacia el progreso científico, renunciando a las tentativas en una dirección metafísica, y operando exclusivamente con nociones que, en principio, corresponden a fenómenos «observables». Si analizamos lo que realmente queremos dar a entender mediante las palabras «aproximación continua», o cómo debemos proceder para verificarla en un caso especial, nos veremos obligados a aceptar una definición tal como la de Cauchy. Esta definición es *estática:* no presupone la idea intuitiva de movimiento. Por el contrario, sólo mediante una definición estática semejante es posible realizar un análisis matemático preciso de la continuidad del movimiento en el tiempo, y dar de lado las paradojas de Zenón en lo que a la ciencia matemática respecta.

En la definición (ε, δ), la variable independiente no se mueve; en ningún sentido físico puede decirse que «tienda» o «se aproxime» al límite x_1. Permanecen todavía estas frases y el símbolo $\rightarrow$, y ningún matemático precisa perder de vista el fondo intuitivo que expresan; pero cuando se trata de verificar la existencia de un límite en un proceso científico real, debe aplicarse la definición (ε, δ) exclusivamente. Que esta idea corresponda satisfactoriamente o no a la noción intuitiva «dinámica», es un problema idéntico al de establecer si los axiomas de la geometría procuran una descripción satisfactoria del concepto intuitivo de espacio. Ambas formulaciones prescinden de algo que tiene intuitivamente una existencia real, pero proporcionan un esquema matemático adecuado para expresar nuestro conocimiento de esos conceptos.

Como en el caso del límite de una sucesión, la clave de la definición de Cauchy consiste en invertir el orden «natural» en el que se consideran las variables. Primero fijamos nuestra atención sobre un intervalo ε para la variable dependiente, tratando después de determinar otro intervalo conveniente δ para la variable independiente. Afirmar que «$f(x) \rightarrow a$, cuando $x \rightarrow x_1$» es sólo una manera breve de decir que eso puede hacerse para cualquier número positivo ε. En

particular, ninguna *parte* de esta afirmación (p. ej., «$x \to x_1$») tiene significado por sí misma.

Es necesario insistir sobre este punto: al hacer «tender» x a x_1 podemos permitir que x sea mayor o menor que x_1, pero excluímos expresamente la igualdad, requiriendo que $x \neq x_1$; x tiende a x_1, pero sin *tomar* nunca el valor x_1. Así podemos aplicar nuestra definición a funciones que no están definidas para $x = x_1$, pero que tienen límites determinados, cuando x tiende a x_1, como, p. ej., la función $f(x) = (x + x^3)/x$, considerada en la página 315. La exclusión de $x = x_1$ corresponde a lo que hacíamos en los límites de las sucesiones a_n para $n \to \infty$; p. ej., $a_n = 1/n$, en la cual no sustituímos n por ∞ en la fórmula.

Sin embargo, cuando x tiende a x_1, $f(x)$ puede aproximarse al límite a de tal manera que existan valores $x \neq x_1$ para los cuales $f(x) = a$; p. ej., al considerar la función $f(x) = x/x$ cuando x tiende a cero, no permitimos nunca que x sea igual a 0, pero $f(x) = 1$, para todo $x \neq 0$, y el límite a existe y es igual a 1, de acuerdo con nuestra definición.

3. El límite de (sen x)/x.—Si x designa la medida de un ángulo en radianes, la expresión (sen x)/x está definida para todo x, excepto $x=0$, para el cual se convierte en el símbolo sin significado 0/0. El lector que posea una tabla de líneas trigonométricas podrá calcular el valor de sen x/x para valores pequeños de x. Como estas tablas proporcionan generalmente las líneas trigonométricas en función del arco expresado en grados, bastará recordar (pág. 289) que la medida en grados x está ligada con la medida en radianes mediante la relación $x = \pi/180\ y = 0{,}01745\ y$, con cinco cifras decimales. De una tabla que proporcione las líneas trigonométricas con cuatro cifras, se deduce:

10°	$x = 0{,}1745$	sen $x = 0{,}1736$	$\dfrac{\text{sen } x}{x} = 0{,}9948$
5°	0,0873	0,0872	0,9988
2°	0,0349	0,0349	1,0000
1°	0,0175	0,0175	1,0000

Aunque, según se ha dicho, estos valores únicamente tienen cuatro cifras decimales exactas, resulta que

$$\text{sen } x/x \to 1 \qquad \text{cuando} \qquad x \to 0. \qquad [1]$$

Daremos ahora una demostración rigurosa de esta relación límite. De la definición de las líneas trigonométricas en el círculo unidad, se deduce que si x es la medida en radianes del ángulo BOC, para $0 < x < \pi/2$, se tiene:

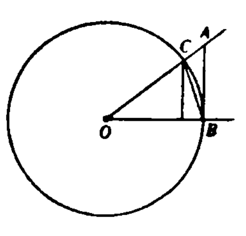


FIG. 169.

Área del triángulo $OBC = {}^1/_2 \cdot 1 \cdot \text{sen}\, x$.

Área del sector circular $OBC = {}^1/_2\, x$ (véase página 290).

Área del triángulo $OBA = {}^1/_2 \cdot 1 \cdot \text{tg}\, x$.

Por tanto, $\text{sen}\, x < x < \text{tg}\, x$.

Dividiendo por sen x, resulta:

$$1 < \frac{x}{\text{sen}\, x} < \frac{1}{\cos x},$$

o

$$\cos x < \frac{\text{sen}\, x}{x} < 1. \qquad [2]$$

Pero $1 - \cos x = (1 - \cos x)\dfrac{1+\cos x}{1+\cos x} = (1-\cos^2 x)/(1+\cos x) =$
$= \text{sen}^2 x/(1+\cos x) < \text{sen}^2 x$, y como $\text{sen}\, x < x$, resulta que

$$1 - \cos x < x^2, \qquad [3]$$

o

$$1 - x^2 < \cos x.$$

Teniendo en cuenta [2], se obtiene definitivamente:

$$1 - x^2 < \frac{\text{sen}\, x}{x} < 1. \qquad [4]$$

Aunque hemos supuesto que $0 < x < \pi/2$, esta desigualdad es también válida para $-\pi/2 < x < 0$, ya que $\dfrac{\text{sen}\,(-x)}{(-x)} = \dfrac{-\text{sen}\, x}{-x} =$
$= \dfrac{\text{sen}\, x}{x}$, y $(-x)^2 = x^2$.

De [4] se deduce inmediatamente la relación [1], pues la diferencia entre $(\text{sen}\, x)/x$ y 1 es menor que x^2, que puede hacerse menor que cualquier número positivo ε, por pequeño que sea, sin más que tomar $|x| < \delta = \sqrt{\varepsilon}$.

Ejercicios:

1. De la desigualdad [3] dedúzcase la relación límite $(1 - \cos x)/x \to 0$, para $x \to 0$.

Determínense los límites, para $x \to 0$, de las siguientes funciones:

2. $\dfrac{\text{sen}^2 x}{x}$ 3. $\dfrac{\text{sen}\, x}{x(x-1)}$ 4. $\dfrac{\text{tg}\, x}{x}$ 5. $\dfrac{\text{sen}\, ax}{x}$

6. $\dfrac{\text{sen}\, ax}{\text{sen}\, bx}$ 7. $\dfrac{x\, \text{sen}\, x}{1 - \cos x}$ 8. $\dfrac{\text{sen}\, x}{x}$, si x está medido en grados.

9. $\dfrac{1}{x} - \dfrac{1}{\text{tg}\, x}$ 10. $\dfrac{1}{\text{sen}\, x} - \dfrac{1}{\text{tg}\, x}$

4. Límites para $x \to \infty$.—Si la variable x es suficientemente grande, la función $f(x) = 1/x$ se hace arbitrariamente pequeña o «tiende a cero». En efecto, el comportamiento de esta función al crecer x es esencialmente el mismo que el de la sucesión $1/n$ al aumentar n. Daremos la definición general: *La función* f(x) *tiene el límite* a *cuando* x *tiende a infinito,* o *en símbolos*

$$f(x) \to a \quad \text{para} \quad x \to \infty,$$

si en correspondencia con todo número positivo ε, *por pequeño que sea, puede encontrarse otro, también positivo,* K *(que depende de* ε*), tal que* $|f(x) - a| < \varepsilon$, *siempre que sea* $|x| > K$. (Compárese ésta con la definición dada en las páginas 315, 316.)

En el caso de la función $f(x) = 1/x$, para la cual es $a = 0$, basta tomar $K = 1/\varepsilon$, como podrá verificar el lector inmediatamente.

Ejercicios:

1. Demuéstrese que la definición anterior de la afirmación

$$f(x) \to a \quad \text{cuando} \quad x \to \infty$$

equivale a esta otra:

$$f(x) \to a \quad \text{cuando} \quad 1/x \to 0.$$

Demuéstrense las siguientes relaciones límites:

2. $\dfrac{x+1}{x-1} \to 1$ cuando $x \to \infty$.

3. $\dfrac{x^2+x+1}{x^2-x-1} \to 1$ cuando $x \to \infty$.

4. $\dfrac{\operatorname{sen} x}{x} \to 0$ cuando $x \to \infty$.

5. $\dfrac{x+1}{x^2+1} \to 0$ cuando $x \to \infty$.

6. $\dfrac{\operatorname{sen} x}{x + \cos x} \to 0$ cuando $x \to \infty$.

7. $\dfrac{\operatorname{sen} x}{\cos x}$ carece de límite para $x \to \infty$.

8. Defínase: «$f(x) \to \infty$ cuando $x \to \infty$». Dése un ejemplo.

Existe una diferencia entre el caso de una función $f(x)$ y una sucesión a_n. En esta última, n puede tender a infinito sólo aumentando, mientras que para una función, x tiende a infinito positiva o negativamente. Si deseamos restringir nuestra atención al comportamiento de $f(x)$ cuando x toma sólo valores *positivos* muy grandes, podemos reemplazar la condición $|x| > K$ por $x > K$, mientras que para valores negativos muy grandes de x, utilizaremos la condición $x < -K$. Para expresar simbólicamente estos dos métodos de aproximación *unilateral* a infinito, escribiremos

$$x \to +\infty \qquad x \to -\infty,$$

respectivamente.

IV. DEFINICIÓN PRECISA DE CONTINUIDAD

Lo que dijimos en las páginas 294-97 equivale al siguiente criterio para reconocer la continuidad de una función: «Una función $f(x)$ es continua en el punto $x = x_1$ si, cuando x tiende a x_1, $f(x)$ tiene como límite $f(x_1)$». Si analizamos esta definición, se ve que consiste en exigir dos cosas enteramente distintas:

a) Debe existir el límite a de $f(x)$, cuando x tiende a x_1.

b) Este límite a debe ser igual al valor $f(x_1)$.

Si en la definición de límite (pág. 316) hacemos $a = f(x_1)$, la condición de continuidad adquiere la forma siguiente: *la función* f(x) *es*

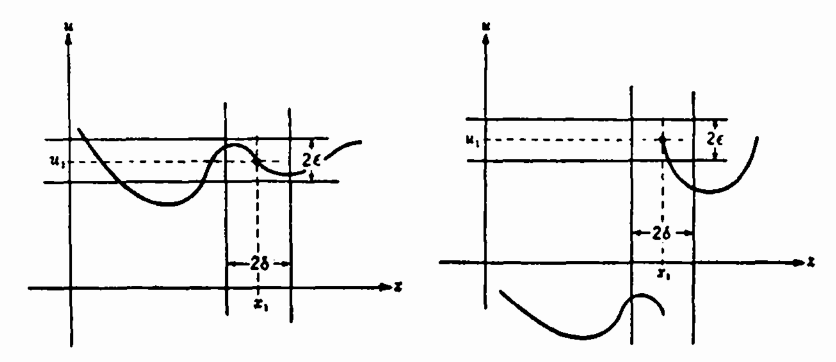


FIG. 170.—Función continua en $x = x_1$. FIG. 171.—Función discontinua en $x = x_1$.

continua para el valor $x = x_1$ *si, para todo número positivo* ε, *por pequeño que sea, puede encontrarse otro* δ, *también positivo (que depende de* ε*), tal que*

$$| f(x) - f(x_1) | < \varepsilon$$

para todo x *que satisfaga la desigualdad*

$$| x - x_1 | < \delta.$$

(La restricción $x \neq x_1$, impuesta en la definición de límite, es innecesaria aquí, puesto que la desigualdad $| f(x) - f(x_1) | < \varepsilon$ se satisface automáticamente.)

Como ejemplo, comprobemos la continuidad de la función $f(x) = x^3$, en el punto $x_1 = 0$. Se tiene

$$f(x_1) = 0^3 = 0.$$

Asignemos ahora a ε cualquier valor positivo, p. ej., $\varepsilon = 1/1000$. Debemos demostrar ahora que si x se mantiene suficientemente pró-

ximo a $x_1 = 0$, los valores correspondientes de $f(x)$ no diferirán de 0 más de 0,001; es decir, estarán situados entre $-0{,}001$ y 0,001. Se ve inmediatamente que no salen de ese intervalo si se impone la condición de que los valores de x difieran de $x_1=0$ en menos de $\delta = \sqrt[3]{0{,}001} = 0{,}1$; pues si $|x| < 0{,}1$, se tiene que $|f(x)| = x^3 < 0{,}001$. De la misma forma, podemos reemplazar $\varepsilon = 0{,}001$ por $\varepsilon = 10^{-4}$, 10^{-5}, o cualquier ofro valor que deseemos. La condición quedará siempre satisfecha si se toma $\delta = \sqrt[3]{\varepsilon}$, ya que si $|x| < \sqrt[3]{\varepsilon}$, $|f(x)| = x^3 < \varepsilon$.

Basándonos en la definición (ε, δ) de la continuidad, es posible demostrar de manera análoga que los polinomios, las funciones racionales y las trigonométricas son funciones continuas, excepto para valores aislados de x, para los cuales la función puede hacerse infinita.

Teniendo en cuenta la gráfica de una función $u = f(x)$, la definición de continuidad toma la siguiente forma geométrica. Elíjase un número positivo ε y trácense paralelas al eje de las x a alturas $f(x_1) - \varepsilon$ y $f(x_1) + \varepsilon$ por encima de él. Entonces, deberá ser posible hallar un número positivo δ, tal que todo el trozo de la gráfica que se encuentra dentro de la banda vertical de anchura 2δ alrededor de x_1 esté contenido dentro de la banda horizontal de anchura 2ε, alrededor de $f(x_1)$. En la figura 170 aparece una función que es continua en x_1, mientras que la figura 171 muestra una función que no lo es. En este último caso, por muy estrecha que sea la banda vertical alrededor de x_1, siempre incluirá una porción de la gráfica que queda fuera de la banda horizontal correspondiente a la elección de ε.

Si afirmo que una función dada $u = f(x)$ es continua para $x = x_1$, quiero decir que estoy dispuesto a convenir el siguiente contrato con el lector. Usted puede elegir cualquier número positivo ε, tan pequeño como desee, pero fijo. Entonces debo determinar otro número δ, también positivo, tal que de $|x - x_1| < \delta$ se deduzca $|f(x) - f(x_1)| < \varepsilon$. No me obligo a ofrecer desde el principio un número δ que sirva para cualquier ε que usted subsiguientemente elija; *mi* elección de δ depende de la *suya* de ε. Si usted puede elegir un solo valor de ε para el cual yo sea incapaz de proporcionar el correspondiente δ, mi aserción queda al descubierto. De ahí que para demostrar que puedo cumplir el contrato, en cualquier caso concreto que se presente de una función $u = f(x)$, generalmente construyo una función positiva explícita

$$\delta = \varphi(\varepsilon)$$

que esté definida para todo número positivo ε, para el cual se pueda demostrar que de $|x - x_1| < \delta$ se deduce siempre $|f(x) - f(x_1)| < \varepsilon$. En el caso de la función $u = f(x) = x^3$, para el valor $x_1 = 0$, la función $\delta = \varphi(\varepsilon)$ era $\delta = \sqrt[3]{\varepsilon}$.

Ejercicios:

1. Demuéstrese que las funciones sen x y cos x son continuas.
2. Demuéstrese la continuidad de $1/(1 + x^4)$ y de $\sqrt{1 + x^2}$.

Parecerá ahora evidente que la definición (ε, δ) de la continuidad coincide con lo que podría considerarse como el conjunto de los hechos observables respecto a una función. En esa forma, está de acuerdo con un principio general de la ciencia moderna según el cual el criterio de la utilidad de un concepto o la «existencia científica» de un fenómeno consiste (al menos en principio) en la posibilidad de su observación, o de su reducción a hechos observables.

V. DOS TEOREMAS FUNDAMENTALES SOBRE LAS FUNCIONES CONTINUAS

1. **Teorema de Bolzano.**—Bernard Bolzano (1781-1848), un sacerdote católico, buen conocedor de la filosofía escolástica, fué uno de los primeros en introducir la moderna idea del rigor en el análisis matemático. Su importante opúsculo *Paradoxien des Unendlichen* apareció en 1850. Por primera vez se reconoció que, si han de utilizarse en toda su generalidad, pueden y deben demostrarse muchas proposiciones aparentemente evidentes, que se refieren a las funciones continuas. Un ejemplo lo constituye el siguiente teorema sobre funciones continuas de una variable. *Una función continua de una variable* x, *positiva para un cierto valor de* x *y negativa para otro, ambos pertenecientes a un intervalo de continuidad cerrado* $a \leqslant x \leqslant b$, *debe tomar el valor cero para algún valor intermedio de* x. Esto es, si $f(x)$ es continua, mientras x varía de a a b, siendo $f(a) < 0$ y $f(b) > 0$, debe existir un valor α de x tal que $a < \alpha < b$ y $f(\alpha) = 0$.

El teorema de Bolzano corresponde perfectamente a nuestra idea intuitiva de una curva continua, la cual debe atravesar al eje x para pasar de un punto situado por debajo de él a otro que se encuentra por encima. La figura 157 pone de manifiesto que esto puede *no* ser cierto para una función discontinua.

*2. **Demostración del teorema de Bolzano.**—Daremos ahora una demostración rigurosa de este teorema. (Siguiendo a Gauss y a otros grandes matemáticos, podríamos aceptar y utilizar ese hecho sin demostración.) Nuestro propósito es el de reducir el teorema a las propiedades fundamentales del sistema numérico real, en particular al postulado de Dedekind-Cantor, el cual se refiere a los encajes de intervalos (pág. 77). Para ello, consideremos el intervalo I, $a \leqslant x \leqslant b$, en el cual está definida $f(x)$, que dividiremos en dos partes por su punto medio $x_1 = (a+b)/2$. Si resultase ser $f(x_1) = 0$, el teorema quedaría demostrado. Sin embargo, si $f(x_1) \neq 0$, $f(x_1)$ será mayor o menor que cero. En cualquiera de los dos casos, cada una de las dos mitades de I gozará ahora de la misma propiedad: $f(x)$ tiene signo distinto en ambos extremos. Llamemos I_1 a este nuevo intervalo y repitamos el mismo proceso, dividiendo I_1 en dos partes iguales; será entonces $f(x) = 0$ en el punto medio de I_1, o, de lo contrario, podremos elegir un intervalo I_2, mitad del I_1,

tal que $f(x)$ tenga signos distintos en ambos extremos. Repitiendo este procedimiento, después de un número finito de bisecciones encontraremos un punto para el cual $f(x) = 0$, o tendremos una sucesión de intervalos encajados $I_1, I_2, I_3, \cdots$ En este último caso, el postulado de Dedekind-Cantor asegura la existencia de un punto α en I, común a todos esos intervalos. Afirmamos que $f(\alpha) = 0$, por lo que α es el punto cuya existencia demuestra el teorema.

Hasta ahora no hemos utilizado la hipótesis inicial de la continuidad de la función. La usaremos ahora para completar la demostración mediante un breve razonamiento indirecto. Demostraremos que $f(\alpha) = 0$, suponiendo lo contrario y deduciendo una contradicción. Supongamos que $f(\alpha) \neq 0$; p. ej., $f(\alpha) = 2\varepsilon > 0$. Dado que $f(x)$ es continua, podremos encontrar un intervalo J (tal vez muy pequeño) de amplitud 2δ, cuyo punto medio sea α y tal que en todo él el valor de

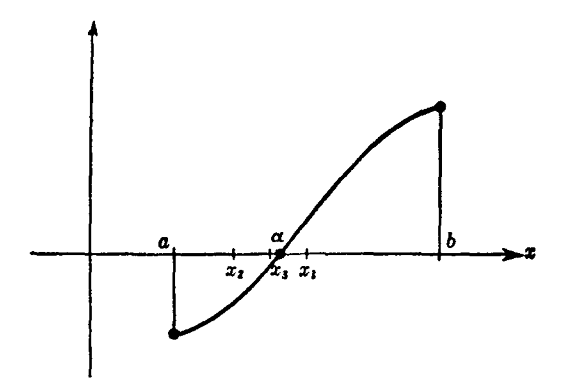


FIG. 172.—Teorema de Bolzano.

$f(x)$ difiera de $f(\alpha)$ en menos de ε. Ahora bien: como $f(\alpha) = 2\varepsilon$, podemos asegurar que $f(x) > \varepsilon$ en todo punto de J, por lo que también se verificará que $f(x) > 0$ en J. Pero como el intervalo J es fijo, bastará tomar n suficientemente grande para que el intervalo I_n quede enteramente dentro de J, ya que la sucesión I_n tiende a 0. Aquí aparece la contradicción, pues de la forma como se eligió I_n, resulta que la función $f(x)$ debe tener signos opuestos en los extremos de todo I_n, por lo que $f(x)$ tiene que tomar valores negativos en algún punto de J. El absurdo a que conduce suponer $f(\alpha) > 0$ o bien $f(\alpha) < 0$, demuestra que es $f(\alpha) = 0$.

3. Teorema de Weierstrass sobre valores extremos.—Karl Weierstrass (1815-1897) formuló otro hecho importante acerca de las funciones continuas, que parece también evidente para la intuición. Tal vez más que a ningún otro matemático se debe a Weierstrass la tendencia moderna hacia el rigor en el análisis matemático. El teorema aludido dice: *Si una función* f(x) *es continua en un intervalo* I, $a \leqslant x \leqslant b$, *incluídos los dos extremos* a *y* b *del intervalo, debe existir por lo menos un punto en* I, *donde* f(x) *alcanza su valor máximo* M, *y otro donde* f(x) *toma su valor mínimo* m. Intuitivamente, esto significa que la gráfica

de la función $u = f(x)$ debe tener, por lo menos, un punto más alto y otro más bajo.

Importa hacer constar que el teorema puede no ser cierto si $f(x)$ deja de ser continua en los extremos del intervalo I; p. ej., la función $f(x) = 1/x$ no tiene un valor máximo en el intervalo $0 < x \leqslant 1$, aunque $f(x)$ es continua en el interior de dicho intervalo. Tampoco precisa una función discontinua tomar un valor máximo o mínimo, aunque esté acotada; p. ej., considérese la función, evidentemente discontinua, definida de la siguiente manera

$$f(x) = x \text{ para } x \text{ irracional,}$$
$$f(x) = {}^1/_2 \text{ para } x \text{ racional,}$$

en el intervalo $0 \leqslant x \leqslant 1$. Esta función toma siempre valores comprendidos entre 0 y 1, y precisamente tan próximos a 0 ó a 1 como queramos, si elegimos un valor irracional de x suficientemente próximo a 0 ó a 1; pero $f(x)$ no puede ser nunca *igual* a 0 ó a 1, puesto que si x es racional, se tiene $f(x) = {}^1/_2$, y para x irracional, resulta $f(x) = x$, por lo que la función nunca alcanza los valores 0 ó 1.

*El teorema de Weierstrass se puede demostrar de modo análogo al seguido para el de Bolzano. Dividimos I en dos semiintervalos cerrados I' e I'', y fijamos nuestra atención sobre I', como el intervalo donde debe buscarse el valor máximo de $f(x)$, *a menos que exista un punto* α *en* I'', *tal que* $f(\alpha)$ *exceda a cualquier valor de* $f(x)$ *en* I'; si así ocurriera, elegiríamos I''. Llamaremos I_1 al intervalo así elegido. Procedemos ahora con I_1 de igual forma que hicimos con I, obteniendo un intervalo I_2, etc. Mediante este proceso, se define una sucesión $I_1, I_2, I_3, \cdots, I_n, \cdots$, de intervalos encajados, que definen un punto z. Demostraremos que el valor $f(z) = M$ es el mayor que toma la función en I; esto es, tal que no puede existir un punto s en I para el cual $f(s) > M$. Supongamos que exista un punto s tal que $f(s) = M + 2\varepsilon$, donde ε es un número positivo (quizá muy pequeño). Alrededor de z como centro, dado que $f(x)$ es continua, limitaremos un pequeño intervalo K, tal que s quede fuera de él y que los valores que en K tome $f(x)$ difieran de $f(z) = M$ en menos de ε, por lo que se tendrá ciertamente, en K, $f(x) < M + \varepsilon$. Pero, para un n suficientemente grande, el intervalo I_n queda dentro de K, y, por otra parte, se definió I_n de tal manera que ningún valor de $f(x)$ para todo x fuera de I_n pudiera superar a los valores de $f(x)$ para los x de I_n. Ya que s está fuera de I_n, y $f(s) > M + \varepsilon$, mientras que en K y, por consiguiente, en I_n, tenemos que $f(x) < M + \varepsilon$, hemos llegado a una contradicción.

De la misma manera puede demostrarse la existencia de un mínimo m, o puede deducirse directamente de lo que ha sido ya demostrado, puesto que el valor mínimo de $f(x)$ es el máximo de $g(x) = -f(x)$.

Puede probarse de manera idéntica el teorema de Weierstrass para funciones continuas de dos o más variables, $x, y, \cdots$ En lugar de un intervalo completo (con sus extremos), debemos considerar un dominio *cerrado*; es decir, un rectángulo en el plano x, y, en el que se incluye su frontera.

Ejercicio: ¿En qué parte de las demostraciones de los teoremas de Bolzano y Weierstrass se utilizó la hipótesis de que $f(x)$ estaba definida y era continua en todo el intervalo *cerrado* $a \leqslant x \leqslant b$, y no simplemente en $a < x \leqslant b$ ó $a < x < b$?

Las demostraciones de los teoremas de Bolzano y Weierstrass tienen un carácter decididamente no-constructivo. No proporcionan un método para encontrar realmente la situación de un cero o del valor máximo o mínimo de una función, con una precisión fijada de antemano, en un número finito de pasos. Se prueba tan sólo la mera existencia, o, mejor dicho, el absurdo de la no-existencia de los valores deseados. Éste es otro ejemplo importante contra el cual han protestado los «intuicionistas» (véase pág. 95); algunos de ellos han insistido en que dichos teoremas deben eliminarse de la matemática. El estudiante de matemática no ha de tomar todo esto más seriamente que la máyor parte de los críticos.

***4. Un teorema sobre sucesiones. Conjuntos compactos.**—Sea x_1, x_2, x_3, ..., una sucesión indefinida de números, distintos entre sí o no, contenidos todos en el *intervalo cerrado* I, $a \leqslant x \leqslant b$. La sucesión puede o no tender a un límite. En todo caso, *es siempre posible extraer otra sucesión* y_1, y_2, y_3, ..., *también indefinida, omitiendo ciertos términos de la dada, y tal que tienda a un límite* y, *contenido en el intervalo* I.

Para demostrar este teorema, dividimos el intervalo I en dos intervalos cerrados I' y I'', mediante el punto medio $(a+b)/2$ de I:

$$I' : a \leqslant x \leqslant \frac{a+b}{2},$$

$$I'' : \frac{a+b}{2} \leqslant x \leqslant b.$$

Por lo menos en uno de estos intervalos, que llamaremos I_1, se encontrarán infinitos términos x_n de la sucesión original. Elijamos cualquiera de ellos, x_n, y llamémosle y_1. Procedamos ahora de la misma manera con el intervalo I_1. Puesto que existen infinitos términos x_n en I_1, deberán existir también infinitos en una al menos de las mitades de I_1, que llamaremos I_2. Por tanto, podremos determinar ciertamente un término x_n de I_2 para el cual $n > n_1$. Elijamos uno entre ellos y llamémosle y_2. Procediendo de esta manera, encontraremos una sucesión de intervalos encajados, I_1, I_2, I_3, ..., y una sucesión parcial y_1, y_2, y_3, ..., de la primitiva, tal que y_n se encuentre en I_n para todo n. Este encaje de intervalos define un punto y de I, y resulta evidente que la sucesión y_1, y_2, y_3, ..., tiene el límite a, como se quería demostrar.

*Estas consideraciones son susceptibles de ser generalizadas en una forma que es típica de la matemática moderna. Consideremos una variable X que se extiende por un conjunto general S, en el cual se ha definido la noción de «distancia» de alguna manera. S puede ser un conjunto de puntos en un plano o en el espacio. Pero esto no es necesario; p. ej., S puede ser el conjunto de todos los triángulos del plano. Si X e Y son dos de esos triángulos, de vértices A, B, C, y A'. B', C', respectivamente, podemos definir la «distancia» entre ambos triángulos mediante el número

$$d(X, Y) = AA' + BB' + CC',$$

donde AA', etc., denota lo que se entiende comúnmente por distancia entre los puntos A y A'. En cualquier conjunto S, en el que esté definida la noción de «distancia», podemos definir también el concepto de sucesión de elementos X_1, X_2, X_3, $\cdots$, que tienden a un elemento límite X de S. Con ello queremos significar que $d(X, X_n) \to 0$, cuando n tiende a ∞.

Diremos ahora que *el conjunto* S *es compacto si de una sucesión cualquiera de elementos* X_1, X_2, X_3, $\cdots$, *de* S, *podemos extraer siempre otra sucesión parcial que tiende a un elemento límite* X *de* S. En el párrafo precedente se ha demostrado que, en este sentido, un intervalo cerrado $a \leq x \leq b$ es un conjunto compacto. De ahí que el concepto de conjunto compacto pueda considerarse como una generalización del de *intervalo cerrado* de la recta numérica. Obsérvese que la recta numérica *en su totalidad* no es un conjunto compacto, puesto que la sucesión de números enteros, 1, 2, 3, 4, 5, $\cdots$, ni tiende a un límite, ni contiene una sucesión parcial convergente. Tampoco un intervalo abierto, sin incluir sus extremos, tal como $0 < x < 1$, es compacto, puesto que la sucesión $1/2$, $1/3$, $1/4$, $\cdots$, o cualquier sucesión parcial de la misma, tiende al límite cero, el cual no es un punto del intervalo abierto. Del mismo modo puede demostrarse que la región del plano formada por los puntos interiores a un cuadrado o rectángulo no es un conjunto compacto, aunque sí lo es cuando se añaden los puntos frontera. Por otra parte, el conjunto de todos los triángulos cuyos vértices son interiores o están sobre la circunferencia de un círculo dado es compacto.

Es también posible generalizar la noción de continuidad al caso en que la variable X varíe en un conjunto cualquiera S, en el cual se ha definido la noción de límite. La función $u = F(X)$, donde u es un número real, se dice que es continua en el elemento X si, para cualquier sucesión de elementos X_1, X_2, $X_3, \cdots$, que converja hacia X, la sucesión correspondiente de números $F(X_1)$, $F(X_2), \cdots$, tiende al límite $F(X)$. (Podría también darse una definición (ε, δ) equivalente.) Es inmediato demostrar que el teorema de Weierstrass se verifica también en el caso general de una función continua definida para los elementos de un conjunto compacto cualquiera:

Si es u = F(X) *una función continua cualquiera, definida en un conjunto compacto* S, *existe siempre un elemento de* S *para el cual* F(X) *alcanza su valor máximo, y también otro en el cual toma su valor mínimo.*

La demostración resulta sencilla una vez asimilados perfectamente los conceptos generales que figuran en el enunciado, pero no proseguiremos en la discusión de este tema. Según se verá en el capítulo VII, el teorema general de Weierstrass es de suma importancia en la teoría de los máximos y mínimos.

VI. ALGUNAS APLICACIONES DEL TEOREMA DE BOLZANO

1. Aplicaciones geométricas.—El sencillo, aunque general, teorema de Bolzano puede utilizarse para demostrar múltiples proposiciones que en modo alguno resultan obvias a primera vista. Comencemos con la siguiente: *Si* A *y* B *son dos áreas del plano, existe siempre una recta del mismo que simultáneamente divide en dos partes iguales a* A *y a* B. Por un «área» queremos significar una porción cualquiera del plano encerrada por un contorno simple cerrado.

Comencemos por elegir un punto fijo P del plano y tracemos por P una semirrecta PR, a partir de la cual mediremos los ángulos. Si tomamos un rayo PS, que forme un ángulo x con PR, existirá en el plano una recta paralela a PS que *dividirá en dos partes iguales el área* A. Pues si consideramos una recta l_1 paralela a PS y situada por completo de un mismo lado de A, trasladándola hasta la posición l_2 (véase Fig. 173), al otro lado de A, la función cuyo valor sea la parte del área A situada a la derecha de la recta, menos el área de A situada a su izquierda, será positiva para la posición l_1 y negativa para la posición l_2. Puesto que esta función es continua, por el teorema de Bolzano debe anularse para alguna posición intermedia l_x, la cual, en consecuencia, dividirá A en dos partes iguales. Para cada valor de x desde $x = 0°$ hasta $x = 360°$, la recta l_x, que divide a A en dos partes iguales, está definida unívocamente.

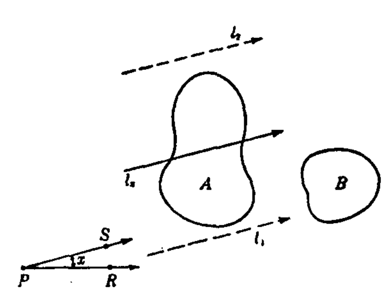


FIG. 173.—División simultánea de dos áreas en partes iguales.

Defínase ahora la función $y = f(x)$ como el área de B a la derecha de l_x, menos el área de B a la izquierda de l_x. Supongamos que la recta l_0, que divide a A en dos partes iguales y tiene la dirección de PR, corta a B de tal manera que deja a la derecha un trozo mayor que a la izquierda. Entonces, para $x = 0°$, y será positiva. Si x aumenta hasta 180°, la recta l_{180}, cuya dirección coincide con la de RP, que divide a A en dos partes iguales, es la misma que l_0; pero tiene sentido contrario, por estar intercambiadas la izquierda y la derecha, de donde se deduce que, numéricamente, el valor de y para $x = 180°$ es igual que para $x = 0°$, aunque de signo opuesto, es decir, negativo. Dado que y es una función continua de x, mientras l_x efectúa una

vuelta completa existe algún valor α de x comprendido entre 0° y 180°, para el cual y se anula. Sigue de ello que la recta l_α divide a A y B simultáneamente en partes iguales, como se quería demostrar.

Obsérvese que, aunque hemos demostrado la *existencia* de una recta que goza de la propiedad enunciada, no hemos dado ningún procedimiento definido para *construirla*. Aparece una vez más este rasgo característico de las demostraciones matemáticas de existencia, en oposición a las construcciones.

Un problema análogo es el siguiente: dado un dominio plano, se desea cortarlo en *cuatro* partes iguales, mediante dos rectas *perpendiculares*. Para demostrar que esto es siempre posible, volvamos al

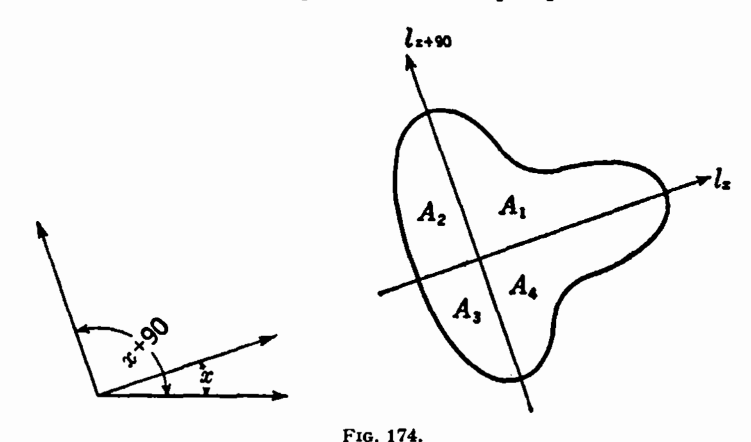


FIG. 174.

problema anterior y precisamente a aquella parte de la demostración donde definimos l_x para cualquier ángulo x; pero prescindamos por completo del área B. En lugar de esto, tracemos la recta l_{x+90}, que es perpendicular a l_x y que también divide a A en dos partes iguales. Si numeramos las cuatro partes de A, como indica la figura 174, tendremos

$$A_1 + A_2 = A_3 + A_4$$

y

$$A_2 + A_3 = A_1 + A_4,$$

de donde, restando la segunda igualdad de la primera, resulta:

$$A_1 - A_3 = A_3 - A_1;$$

es decir,

$$A_1 = A_3,$$

y, por tanto,

$$A_2 = A_4.$$

Así, pues, si podemos demostrar que existe un ángulo α tal que para l_α sea

$$A_1(\alpha) = A_2(\alpha),$$

nuestro teorema quedará demostrado, ya que para tal ángulo las cuatro partes serán iguales. Con este propósito, definimos una función $y = f(x)$ trazando una recta l_x y haciendo $f(x) = A_1(x) - A_2(x)$.

Para $x = 0^\circ$, $f(0) = A_1(0) - A_2(0)$ puede ser positiva. En ese caso, para $x = 90^\circ$, $f(90) = A_1(90) - A_2(90) = A_2(0) - A_3(0) = A_2(0) - A_1(0)$, será negativa. En consecuencia, ya que $f(x)$ varía de una manera continua, cuando x pasa de 0° a 90° existirá un cierto valor α entre 0° y 90°, para el cual $f(\alpha) = A_1(\alpha) - A_2(\alpha) = 0$. Entonces, la recta l_α y su perpendicular $l_{\alpha+90}$ dividirán al dominio en cuatro partes iguales.

Es interesante observar que estos problemas se pueden generalizar al caso de tres o más dimensiones. En tres dimensiones, el problema se enuncia así: Dados tres sólidos en el espacio, encontrar un plano que los divida simultáneamente en partes iguales. Se puede demostrar, partiendo del teorema de Bolzano, que esto es siempre posible. Para más de tres dimensiones el teorema sigue siendo válido, pero la demostración requiere recursos de orden más elevado.

***2. Aplicación a un problema de mecánica.**—Terminaremos este capítulo discutiendo un problema de mecánica, aparentemente difícil, pero que es sencillo de resolver por un razonamiento basado en los conceptos de continuidad. (Este problema fué sugerido por H. Whitney.)

Supongamos un tren que recorre un trayecto rectilíneo de vía entre las estaciones A y B. No es preciso que el móvil tenga velocidad o aceleración uniformes durante el recorrido; el tren puede acelerarse, retardarse, detenerse y aun retroceder durante algún tiempo antes de alcanzar B. Pero se supone conocido de antemano el movimiento exacto del tren; es decir, se supone dada la función $s = f(t)$, donde s es la distancia del tren a la estación A y t el tiempo, medido a partir del instante de salida. Sobre el suelo de uno de los vagones hay una varilla conectada de forma que puede moverse sin rozamiento, hacia delante o hacia atrás, hasta tocar el suelo. Si llega a tocar el suelo, supondremos que permanece en esta posición definitivamente. Nos preguntamos si es posible colocar la varilla de tal manera que si se abandona en el momento en que el tren inicia su movimiento, quedando entonces exclusivamente sujeta a la influencia de la gravedad y del movimiento del tren, no caiga al suelo durante todo el trayecto desde A hasta B. Podrá parecer sumamente improbable que, para

cualquier ley de movimiento, dada la acción conjunta de la gravedad y de las fuerzas de reacción, sea siempre posible el mantenimiento de dicho equilibrio con la única condición de elegir convenientemente la posición inicial de la varilla; con todo, nuestra afirmación es que dicha posición existe siempre.

Por paradójica que esta aserción parezca a primera vista, puede demostrarse fácilmente si nos concentramos en su carácter esencialmente topológico. No se precisa para ello de un conocimiento detallado de las leyes de la dinámica y sólo es necesario suponer realizada la siguiente hipótesis de naturaleza física: *El movimiento de la varilla depende con continuidad de su posición inicial.* Caractericemos la posición inicial de la varilla por el ángulo x, que forma inicialmente con el suelo del vagón, y por el ángulo y, que forma con dicho suelo al

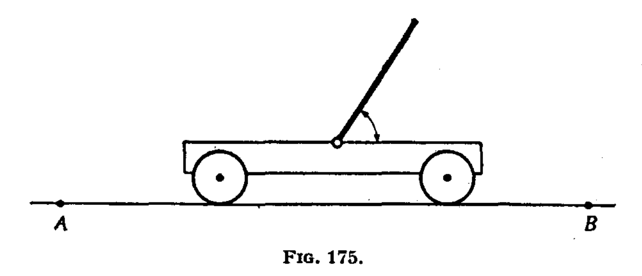


FIG. 175.

final del trayecto, cuando el tren llega al punto B. Si la varilla ha caído al suelo, o bien es $y = 0$ ó $y = \pi$. Para una posición inicial dada x, la posición final y estará, de acuerdo con nuestra hipótesis, unívocamente determinada mediante una función $y = g(x)$, que es continua y toma los valores $y = 0$ para $x = 0$ e $y = \pi$ para $x = \pi$ (esto último expresa, simplemente, que la varilla permanecerá sobre el suelo si ésta es su posición al iniciar el viaje). Recordemos que por ser $g(x)$ una función continua en el intervalo $0 \leqslant x \leqslant \pi$, toma todos los valores entre $g(0) = 0$ y $g(\pi) = \pi$; en consecuencia, para cualquier valor de y, p. ej., para $y = \pi/2$, existe un valor determinado de x tal que $g(x) = y$. En particular, existirá una posición inicial para la cual la posición final de la varilla en B será perpendicular al suelo. (NOTA: No debe olvidarse en este razonamiento que el movimiento del tren está dado de antemano.)

Es evidente que la argumentación es por completo teórica; si el viaje es de larga duración o si la ley de movimiento del tren, expresada por $s = f(t)$, es muy arbitraria, entonces el número de posiciones iniciales x para las que la posición final $g(x)$ es distinta de 0 o de π

será muy pequeño, como sabe cualquiera que haya intentado mantener verticalmente una aguja sobre una superficie plana durante un tiempo apreciable. Con todo, nuestro razonamiento será apreciado incluso por una mentalidad práctica, pues pone de manifiesto la posibilidad de obtener resultados cualitativos en dinámica por simple razonamiento, sin recurrir a una manipulación experimental.

Ejercicios:

1. Utilizando el teorema de la página 326, demuéstrese que es posible generalizar el razonamiento anterior para el caso de un viaje de duración infinita.

2. Generalícese para el caso en que el tren se mueve siguiendo una curva plana cualquiera y la varilla puede caer en cualquier dirección. Obsérvese que no es posible representar un disco circular sobre una circunferencia exclusivamente mediante una transformación continua que deje fijo todo punto de la circunferencia (véase pág. 267).

3. Demuéstrese que el tiempo que precisa la varilla para caer al suelo, si el movimiento del tren es estacionario y la varilla se aparta un ángulo ε de la posición vertical, tiende a infinito al tender ε a cero.

SUPLEMENTO AL CAPÍTULO VI

MÁS EJEMPLOS SOBRE LÍMITES Y CONTINUIDAD

I. EJEMPLOS DE LÍMITES

1. Observaciones generales.—En muchos casos, puede demostrarse la convergencia de una sucesión a_n mediante el recurso de hallar otras dos sucesiones b_n y c_n, cuyos términos tengan una estructura más sencilla que los de la primera y tales que

$$b_n \leqslant a_n \leqslant c_n \qquad [1]$$

para todo n. *Entonces, podemos demostrar que si las sucesiones* b_n *y* c_n *convergen ambas hacia el mismo límite* α, *también* a_n *tiende a dicho límite* α. Queda a cargo del lector la demostración formal de esta afirmación.

Es evidente que al aplicar este procedimiento, deberán utilizarse desigualdades; en consecuencia, conviene recordar algunas reglas elementales que rigen las operaciones aritméticas con desigualdades.

1. Si $a > b$, es $a + c > b + c$. (Puede sumarse cualquier número a los dos miembros de una desigualdad.)
2. Si $a > b$ y el número c es *positivo*, se tiene $ac > bc$. (Puede multiplicarse una desigualdad por cualquier número positivo.)
3. Si $a < b$, es $-b < -a$. (Se invierte el sentido de una desigualdad multiplicándola por -1.) Así, de $2 < 3$, resulta $-3 < -2$.
4. Si a y b tienen el mismo signo, y es $a < b$, resulta $1/a > 1/b$.
5. $|a + b| \leqslant |a| + |b|$.

2. Límite de q^n.—Si q es un número mayor que 1, la sucesión q^n llegará a exceder a cualquier número, como lo hace la siguiente: 2, 2^2, 2^3, ..., en la cual es $q = 2$; la sucesión «tiende a infinito» (véase página 305). En el caso general, la demostración se basa en la importante desigualdad (demostrada en la pág. 22)

$$(1 + h)^n \geqslant 1 + nh > nh, \qquad [2]$$

donde h es un número positivo cualquiera. Hagamos $q = 1 + h$, siendo $h > 0$. Se tiene entonces

$$q^n = (1 + h)^n > nh.$$

Si k es un número positivo cualquiera, por grande que sea, para todo $n > k/h$ se deduce que

$$q^n > nh > k,$$

de donde resulta que $q^n \to \infty$.

Si $q = 1$, todos los términos de la sucesión q^n serán iguales a 1, y éste es también el límite de la sucesión. Si q es negativo, q^n tomará alternativamente valores positivos y negativos, y no tendrá límite si $q \leqslant -1$.

Ejercicio: Dése una demostración rigurosa de esta última afirmación.

En la página 73 demostramos que $q^n \to 0$ si es $-1 < q < 1$. Podemos dar otra demostración muy sencilla de ese hecho. Consideraremos primero el caso $0 < q < 1$; los números q, q^2, q^3, ... forman una sucesión monótona decreciente, acotada inferiormente por 0. De acuerdo con lo dicho en la página 306, la sucesión debe tender a un límite: $q^n \to a$. Multiplicando ambos miembros de esa relación por q se tiene: $q^{n+1} \to aq$.

Pero q^{n+1} debe tener el mismo límite, ya que no importa que el exponente sea n o $n + 1$; luego $aq = a$, o sea $a(q - 1) = 0$. Puesto que $1 - q \neq 0$ esa igualdad implica que $a = 0$.

Si $q = 0$, la afirmación $q^n \to 0$ es trivial. Si $-1 < q < 0$, se tiene $0 < |q| < 1$, de donde se sigue que $|q^n| = |q|^n \to 0$, de acuerdo con el razonamiento anterior. De todo ello resulta que se tiene siempre $q^n \to 0$ para $|q| < 1$, con lo que queda completada la demostración.

Ejercicios: Demuéstrese que para $n \to \infty$:
1. $[x^2/(1 + x^2)]^n \to 0$.
2. $[x/(1 + x^2)]^n \to 0$.
3. $[x^3/(4 + x^2)]^n$ tiende a infinito para $x > 2$ y a cero para $|x| < 2$.

3. Límite de $\sqrt[n]{p}$.—La sucesión $a_n = \sqrt[n]{p}$, es decir, la sucesión p, $\sqrt{p}$, $\sqrt[3]{p}$, $\sqrt[4]{p}$, ... tiene el límite 1 para todo número positivo fijo p:

$$\sqrt[n]{p} \to 1 \quad \text{cuando} \quad n \to \infty. \qquad [3]$$

(Mediante el símbolo $\sqrt[n]{p}$ representamos, como de costumbre, la raíz enésima positiva. Para los números p negativos no existen raíces enésimas reales, si n es par.)

Para demostrar la relación [3], comencemos por suponer que $p > 1$; entonces, $\sqrt[n]{p}$ será también mayor que 1, por lo que podemos escribir:

$$\sqrt[n]{p} = 1 + h_n,$$

donde h_n es un número positivo que depende de n. De acuerdo con la desigualdad [2] resulta que

$$p = (1 + h_n)^n > nh_n .$$

Al dividir por n, vemos que

$$0 < h_n < p/n .$$

Puesto que las sucesiones $b_n = 0$ y $c_n = p/n$ tienen ambas el límite 0, de lo dicho en la página 333 se sigue que h_n también tiene límite 0, con lo que nuestra afirmación queda demostrada para $p > 1$. Éste es un ejemplo típico de un método de determinación de límites, que consiste en encerrar la sucesión problema entre otras dos, cuyos límites son conocidos o más fáciles de obtener.

Incidentalmente, hemos deducido una estimación de la diferencia h_n, entre $\sqrt[n]{p}$ y 1, encontrando que ésta es siempre menor que p/n.

Si $0 < p < 1$, entonces $\sqrt[n]{p} < 1$, por lo que podemos escribir:

$$\sqrt[n]{p} = \frac{1}{1 + h_n},$$

donde h_n representa de nuevo un número positivo que depende de n. De ello se deduce que

$$p = \frac{1}{(1 + h_n)^n} < \frac{1}{nh_n};$$

esto es,

$$0 < h_n < \frac{1}{np}$$

De esta última relación concluímos que h_n tiende a cero al crecer n; pero, dado que $\sqrt[n]{p} = 1/(1 + h_n)$, resulta que $\sqrt[n]{p} \to 1$.

El efecto nivelador de la extracción de la raíz enésima, que hace tender hacia 1 a cualquier número, al aumentar n, es incluso eficaz aun cuando la cantidad bajo el signo radical no permanezca constante. Vamos a demostrar que la sucesión: 1, $\sqrt{2}$, $\sqrt[3]{3}$, $\sqrt[4]{4}$, ..., tiende a 1; es decir, que

$$\sqrt[n]{n} \to 1$$

al tender n a infinito. Es posible también deducir este caso por aplicación de la desigualdad [2]. En lugar de tomar la raíz n-ésima de n, tomemos la raíz n-ésima de $\sqrt{n}$. Si escribimos $\sqrt[n]{\sqrt{n}} = 1 + k_n$,

siendo k_n un número positivo que depende de n, esta desigualdad nos proporciona $\sqrt{n} = (1 + k_n)^n > nk_n$; por lo que

$$k_n < \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}}$$

De donde resulta

$$1 < \sqrt[n]{n} = (1 + k_n)^2 = 1 + 2k_n + k_n^2 < 1 + \frac{2}{\sqrt{n}} + \frac{1}{n}$$

El segundo miembro de esta desigualdad tiende a 1, al crecer n, por lo que $\sqrt[n]{n}$ debe tender también a 1.

4. Las funciones discontinuas como límites de funciones continuas.—Podemos considerar límites de sucesiones a_n cuyos términos no sean constantes, sino dependientes de una variable x; p. ej., $a_n = f_n(x)$. Si esta sucesión converge cuando $n \to \infty$, el límite será a su vez una función de x,

$$f(x) = \text{lím}\, f_n(x).$$

Estas representaciones de ciertas funciones $f(x)$ como límites de otras son, a menudo, muy útiles para reducir funciones «superiores» $f(x)$ a las funciones elementales $f_n(x)$.

Esto es cierto, en particular, en la representación de las funciones discontinuas mediante fórmulas explícitas. Consideremos, p. ej., la sucesión $f_n(x) = 1/(1 + x^{2n})$. Para $|x| = 1$, tenemos $x^{2n} = 1$ y, en consecuencia, $f_n(x) = {}^1/_2$, cualquiera que sea n, de donde resulta $f_n(x) \to {}^1/_2$. Para $|x| < 1$, se tiene que $x^{2n} \to 0$, o sea que $f_n(x) \to 1$, mientras que para $|x| > 1$ tenemos $x^{2n} \to \infty$, de donde se deduce que $f_n(x) \to 0$. En resumen,

$$f(x) = \text{lím}\, \frac{1}{1 + x^{2n}} = \begin{cases} 1 & \text{para } |x| < 1, \\ 1/2 & \text{para } |x| = 1, \\ 0 & \text{para } |x| > 1. \end{cases}$$

En este caso la función discontinua $f(x)$ está representada como el límite de una sucesión de funciones racionales continuas.

Otro ejemplo interesante de carácter análogo es el de la sucesión

$$f_n(x) = x^2 + \frac{x^2}{1 + x^2} + \frac{x^2}{(1 + x^2)^2} + \cdots + \frac{x^2}{(1 + x^2)^n}$$

Para $x = 0$ todos los valores de $f_n(x)$ son cero, luego $f(0) = \text{lím}\, f_n(0) = 0$. Para $x \neq 0$, la expresión $1/(1 + x^2) = q$ es positiva y menor que 1; las propiedades conocidas de la serie geométrica nos permiten asegu-

rar la convergencia de $f_n(x)$ para $n \to \infty$. El límite, es decir, la suma de la progresión geométrica indefinidamente prolongada es: $\frac{x^2}{1-q} = \frac{x^2}{1 - \frac{1}{1+x^2}} = 1 + x^2$. Se ve así que $f_n(x)$ tiende a la función $f(x) = 1 + x^2$ para $x \neq 0$ y a $f(x) = 0$ para $x = 0$. Esta función tiene una discontinuidad evitable para $x = 0$.

***5. Límites por iteración.**—Con frecuencia, los términos de una sucesión son de tal naturaleza que se deduce a_{n+1} de a_n mediante el mismo procedimiento por el cual se calculó a_n a partir de a_{n-1}; repitiendo el proceso indefinidamente, se obtiene toda la sucesión partiendo de un término inicial dado. En tales casos, decimos que se trata de un proceso de «iteración».

Por ejemplo, la sucesión

$$1, \sqrt{1+1}, \sqrt{1+\sqrt{2}}, \sqrt{1+\sqrt{1+\sqrt{2}}}, \cdots$$

tiene una ley de formación de ese tipo; cada término posterior al primero se forma extrayendo la raíz cuadrada de 1 más el que le precede. Así, la fórmula

$$a_1 = 1, \quad a_{n+1} = \sqrt{1 + a_n}$$

define toda la sucesión. Calculemos su límite: evidentemente, a_n es mayor que 1 para $n > 1$; además, a_n es una sucesión monótona creciente, pues

$$a_{n+1}^2 - a_n^2 = (1 + a_n) - (1 + a_{n-1}) = a_n - a_{n-1}.$$

De aquí resulta que si es $a_n > a_{n-1}$, será también $a_{n+1} > a_n$; pero sabemos que $a_2 - a_1 = \sqrt{2} - 1 > 0$, de donde deducimos por inducción que $a_{n+1} > a_n$ para todo n; es decir, que la sucesión es monótona creciente. Además, está acotada, pues de lo anterior resulta:

$$a_{n+1} = \frac{1 + a_n}{a_{n+1}} < \frac{1 + a_{n+1}}{a_{n+1}} = 1 + \frac{1}{a_{n+1}} < 2.$$

Por el principio de las sucesiones monótonas deducimos que, para $n \to \infty$, $a_n \to a$, siendo a un número comprendido entre 2 y 1. Es fácil ver que a es la raíz positiva de la ecuación $x^2 = 1 + x$, pues, para $n \to \infty$, la ecuación $a_{n+1}^2 = 1 + a_n$ se convierte en $a^2 = 1 + a$. Resolviéndola, se encuentra que la raíz positiva es $a = (1 + \sqrt{5})/2$. Podríamos resolver esta ecuación mediante un proceso de iteración,

lo que nos proporcionaría el valor de la raíz con cualquier grado de aproximación deseado, siempre que se avanzase lo suficiente en dicho proceso.

Es posible resolver muchas otras ecuaciones algebraicas por procedimientos análogos de iteración; p. ej., podemos escribir la ecuación cúbica $x^3 - 3x + 1 = 0$ de la siguiente manera:

$$x = \frac{1}{3 - x^2}$$

Si elegimos ahora un valor cualquiera para a_1, p. ej., $a_1 = 0$, y ponemos:

$$a_{n+1} = \frac{1}{3 - a_n^2},$$

obtendremos la sucesión: $a_2 = 1/3 = 0{,}3333...$; $a_3 = 9/26 = 0{,}3461$ ...; $a_4 = 676/1947 = 0{,}3472$ ..., etc. Puede demostrarse que la sucesión a_n, obtenida de esta forma, converge hacia el límite $a = 0{,}3473...$, que es una solución de la ecuación cúbica dada. Estos procesos de iteración tienen una gran importancia, tanto en la matemática pura, donde proporcionan «demostraciones de existencia», como en la aplicada, en cuyo caso suministran métodos aproximados para la resolución de muy distintas clases de problemas.

Ejercicios sobre límites. Para $n \to \infty$:

1. Demuéstrese que $\sqrt{n+1} - \sqrt{n} \to 0$. (Escríbase la diferencia en la forma

$$\frac{\sqrt{n+1} - \sqrt{n}}{\sqrt{n+1} + \sqrt{n}} \cdot (\sqrt{n+1} + \sqrt{n}).)$$

2. Determínese el límite de $\sqrt{n^2 + a} - \sqrt{n^2 + b}$.
3. Hállese el límite de $\sqrt{n^2 + an + b} - n$.
4. Calcúlese el límite de $\dfrac{1}{\sqrt{n+1} + \sqrt{n}}$
5. Demuéstrese que el límite de $\sqrt[n]{n+1}$ es 1.
6. ¿Cuál es el límite de $\sqrt[n]{a^n + b^n}$ si $a > b > 0$?
7. ¿Cuál es el límite de $\sqrt[n]{a^n + b^n + c^n}$ si $a > b > c > 0$?
8. Más adelante veremos (pág. 459) que $e = \lim\,[1 + (1/n)]^n$. ¿Cuál será entonces el límite de $(1 + 1/n^2)^n$?

II. UN EJEMPLO SOBRE CONTINUIDAD

La demostración rigurosa de la continuidad de una función requiere la comprobación explícita de la definición dada en la página 321. A veces esto entraña un procedimiento muy lento, si bien, afortunadamente

como veremos en el capítulo VIII, la continuidad es una consecuencia de la diferenciabilidad. Ya que esta última propiedad se establecerá sistemáticamente para todas las funciones elementales, podemos seguir el procedimiento habitual de omitir las demostraciones enojosas de la continuidad en cada caso particular. Pero, como un ejemplo aclaratorio de la definición general, analizaremos otro caso: la función $f(x) = 1/(1 + x^2)$. Podemos restringir la variación de x a un intervalo prefijado $|x| \leqslant M$, siendo M un número arbitrariamente elegido. Si escribimos

$$f(x_1) - f(x) = \frac{1}{1 + x_1^2} - \frac{1}{1 + x^2} = \frac{x^2 - x_1^2}{(1 + x^2)(1 + x_1^2)} =$$

$$= (x - x_1)\frac{(x + x_1)}{(1 + x^2)(1 + x_1^2)},$$

encontramos que para $|x| \leqslant M$ y $|x_1| \leqslant M$

$$|f(x_1) - f(x)| < |x - x_1||x + x_1| < |x - x_1| \cdot 2M,$$

de donde se deduce claramente que el primer miembro será menor que cualquier número positivo ε, con tal que $|x_1 - x| < \delta = \varepsilon/2M$.

Deberá observarse que hemos sido muy generosos en nuestras apreciaciones. Para valores mayores de x y x_1, el lector podrá ver por sí mismo que hubiera bastado tomar un δ mucho mayor.

CAPÍTULO VII

MÁXIMOS Y MÍNIMOS

Introducción.—Un segmento rectilíneo es la línea más corta entre dos puntos. Un arco de círculo máximo es la curva más corta que une dos puntos de una superficie esférica. Entre todas las curvas planas cerradas de la misma longitud, la circunferencia encierra el área mayor; entre todas las superficies cerradas de igual área, la esfera encierra el mayor volumen.

Estas propiedades de máximo y mínimo eran conocidas de los griegos, aunque a menudo se enunciaban los resultados sin serio intento de dar la demostración. Se atribuye a Herón de Alejandría (siglo I a. de J.C.), uno de los más significativos descubrimientos de la ciencia griega. Se sabía, desde mucho tiempo antes, que un rayo de luz procedente de un punto P y que incide sobre un espejo plano L, en un punto R, se refleja en dirección a un punto Q, tal que PR y QR forman ángulos iguales con el espejo. Herón encontró que si R' es otro punto del espejo, la distancia total $PR' + R'Q$ es mayor que la distancia $PR + QR$. Este teorema, que demostraremos a continuación, caracteriza la trayectoria real PQR, de la luz, como el camino más corto posible entre P y Q, que toca al espejo, descubrimiento que puede considerarse como el germen de la óptica geométrica.

Es muy natural que los matemáticos se interesen por cuestiones de este tipo. En la vida diaria se presentan constantemente problemas de máximos y mínimos, de lo «óptimo» y lo «peor». Muchas cuestiones de importancia práctica se plantean de este modo; p. ej., ¿qué forma ha de tener una nave para presentar una resistencia mínima al agua? ¿Cuál es el recipiente cilíndrico, construído con una cantidad dada de chapa, que tiene un volumen máximo?

Habiéndose iniciado en el siglo XVII, la teoría general de los valores extremos—máximos y mínimos—se ha convertido en uno de los principios sistemáticos integrantes de la ciencia. Los primeros pasos de Fermat en su cálculo diferencial estuvieron animados por el deseo de estudiar las cuestiones de máximos y mínimos por medio de métodos generales. En el siglo siguiente, se amplió el campo de aplicación de estos métodos mediante la invención del *cálculo de variaciones*. Era cada vez más evidente que las leyes físicas de la Naturaleza encontraban su expresión más adecuada en el principio de mínimo, el cual

proporciona un acceso natural a una solución más o menos completa de problemas particulares. Una de las realizaciones más notables de la matemática contemporánea es la teoría de los valores estacionarios, generalización del concepto de valores extremos, que combina el análisis con la topología. La exposición que haremos de toda esta cuestión será completamente elemental.

I. PROBLEMAS DE GEOMETRÍA ELEMENTAL

1. Triángulo de área máxima, dados dos lados.—Se dan dos segmentos a y b, y se pide hallar el triángulo de área máxima que tenga esos segmentos por lados. La solución es sencillamente el triángulo rectángulo de catetos a y b. Pues si consideramos cualquier triángulo (Fig. 176), dos de cuyos lados sean a y b, y h la altura respecto a a como base, el área del triángulo será $A = {}^1/_2ah$. Ahora bien: ${}^1/_2ah$ será máximo cuando sea máximo h, lo que ocurrirá si h coincide con b; es decir, para un triángulo rectángulo. Por tanto, el área máxima es ${}^1/_2ab$.

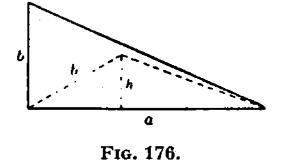


FIG. 176.

2. Teorema de Herón. Propiedad extremal de los rayos luminosos. Sea una recta L y dos puntos P y Q al mismo lado de L; ¿para qué punto R de L es $PR + RQ$ la trayectoria mínima de P a Q tocando a L? Éste es el problema de Herón del rayo luminoso. (Si L fuera la ribera de un río y alguien tuviera que ir de P a Q tan rápidamente como fuera posible, teniendo que sacar durante el camino un cubo de agua de L, tendría que resolver precisamente este problema.)

Para hallar la solución, reflejemos P sobre L como si fuera un espejo, obteniendo el punto P', de tal manera que L es la mediatriz de PP'. La recta $P'Q$ corta a la L en el punto buscado R. Es inmediata la demostración de que $PR + RQ$ es menor que $PR' + R'Q$, para cualquier otro punto R' de L. Puesto que $PR = P'R$ y $PR' = P'R'$, se deduce que $PR + RQ = P'R + RQ = P'Q$, y que $PR' + R'Q = P'R' + R'Q$. Pero como $P'R' + R'Q$ es mayor que $P'Q$ (puesto que la suma de dos lados cualesquiera de un triángulo es mayor que el tercero), resulta que $PR' + R'Q$ es mayor que $PR + RQ$, como se quería probar. En lo que sigue, supondremos que ni P ni Q se encuentran sobre L.

De la figura 177 se deduce que son iguales los ángulos 2 y 3, así como los 2 y 1, luego también son iguales 1 y 3. En otras palabras, R *es un punto tal que* PR *y* QR *forman ángulos iguales con* L. De esto

se deduce que un rayo luminoso que se refleje en L (del cual se sabe por experiencia que forma ángulos iguales de incidencia y de reflexión) sigue en realidad la trayectoria más corta que conduce de P a Q pasando por L, como antes se afirmó en la introducción.

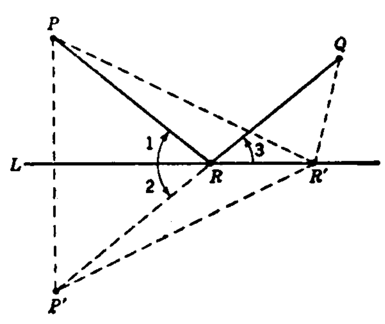


FIG. 177.—Teorema de Herón.

Puede generalizarse el problema para incluir el caso de varias rectas, L, M, ... Sean, p. ej., dos rectas L y M, y dos puntos P y Q, situados como en la figura 178; el problema consiste en encontrar la trayectoria mínima de P a L, de aquí a M y después a Q. Sea Q' el simétrico de Q, respecto a la recta M, y Q'' el de Q' respecto a L. Trácense las rectas PQ'', que corta a L en R, y RQ', que corta a M en S; los puntos R y S son los buscados, siendo $PR + RS + SQ$ el camino mínimo de P a Q, que pasa por L y M. La demostración es muy parecida a la del problema anterior y queda al cuidado del lector. Si L y M fueran espejos, un rayo

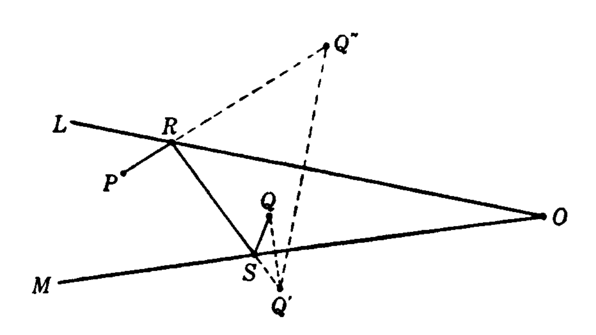


FIG. 178.—Reflexión sobre dos espejos.

luminoso procedente de P, que se reflejara en L y después en M, para seguir hacia Q, incidiría sobre L en R y sobre M en S, de donde resulta otra vez que el rayo luminoso sigue la trayectoria de longitud mínima.

Cabe preguntarse cuál es la trayectoria mínima de P a M, después a L y desde allí hasta Q. Esto daría una trayectoria $PRSQ$ (Fig. 179), determinada de manera análoga a la anterior $PRSQ$. La

longitud de la primera trayectoria puede ser mayor, igual o menor que la de la segunda.

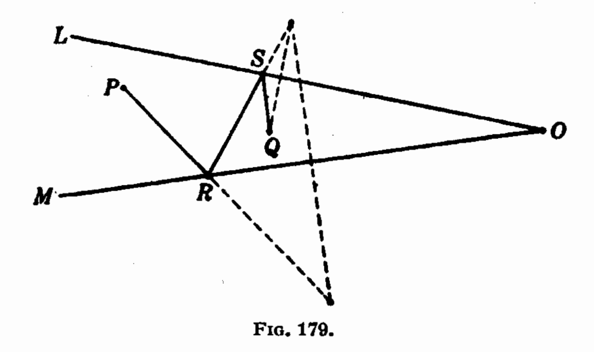


FIG. 179.

***Ejercicio:** Demuéstrese que la primera trayectoria es menor que la segunda si O y R están a un mismo lado de la recta PQ. ¿En qué caso tendrán ambas trayectorias igual longitud?

3. Aplicaciones a problemas sobre triángulos.—Con ayuda del teorema de Herón puede obtenerse fácilmente la solución de los dos problemas siguientes: *a)* Dados el área A y un lado $c = PQ$ de un triángulo, determínese éste de forma que la suma de los otros dos lados a y b sea mínima. Conocer el lado c y el área A de un triángulo equivale a conocer el lado c y la altura correspondiente h, ya que $A = {}^1/_2hc$. Con referencia a la figura 180 el problema consiste, por tanto, en determinar un punto R tal que la distancia del mismo a la recta PQ sea igual a la altura dada h, y de manera que la suma $a + b$ sea mínima. De la primera condición se deduce que R debe ser un punto de la paralela a PQ a la distancia h. La solución nos la procura el teorema de Herón para el caso especial en que P y Q equidistan de L; el triángulo pedido, PRQ, es isósceles.

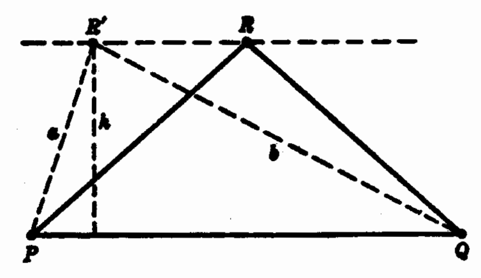


FIG. 180.—Triángulo del perímetro mínimo, dadas su área y su base.

b) De un triángulo se conocen el lado c y la suma $a + b$ de los otros dos lados; de todos los triángulos posibles, hállese el de área máxima. Éste no es otro que el problema inverso del *a)*, y la solución es nuevamente el triángulo isósceles para el cual $a = b$. Según

acabamos de demostrar, en este triángulo es mínimo el valor de $a + b$; es decir, en cualquier otro triángulo de base c y de igual área, es mayor el valor de $a + b$. Además, resulta evidente de *a)* que cualquier triángulo de base c y área mayor que la del triángulo isósceles tiene también mayor la suma de a y b. En consecuencia, cualquier otro triángulo que tenga los mismos valores para $a + b$ y c debe tener un área menor, de forma que el triángulo isósceles es el de área máxima, dados c y $a + b$.

4. Propiedades de las tangentes a la elipse y a la hipérbola. Propiedades extremales de las mismas.—El problema de Herón está relacionado con algunos teoremas geométricos importantes. Hemos demostrado que si R es un punto de L, tal que $PR + RQ$ es mínima, PR y QR forman ángulos iguales con L. Llamaremos $2a$ a esa distancia total mínima. Designemos por p y q las distancias de un punto cualquiera del plano a P y Q, respectivamente, y consideremos el lugar geométrico de *todos* los puntos del plano, para los cuales, $p + q = 2a$. Este lugar es una elipse, de focos P y Q, y que pasa por el punto R de la recta L. Además, L *debe ser tangente a la elipse en* R. Si L cortase a la elipse en un punto distinto de R, debería existir un segmento de L interior a la elipse, y para cada punto del mismo, $p + q$ sería menor que $2a$, ya que es fácil ver que $p + q$ es menor que $2a$ para los puntos interiores a la elipse, y mayor para todo punto exterior. Puesto que sabemos que la desigualdad $p + q \geqslant 2a$ no se cumple para ningún otro punto de L, esta última hipótesis es imposible. De ahí que L deba ser tangente a la elipse en R. Pero como sabemos que PR y RQ forman ángulos iguales con L, hemos demostrado incidentalmente el siguiente teorema importante: Cualquier tangente a una elipse forma ángulos iguales con los radios vectores que unen el punto de tangencia a los focos.

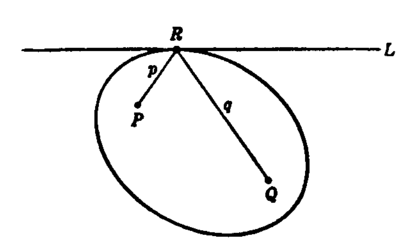


FIG. 181.—Propiedad de la tangente a la elipse.

El siguiente problema está íntimamente relacionado con la conclusión anterior: Dada una recta L y dos puntos P y Q, situados en lados *opuestos* de L (Fig. 182), hállese un punto R de L, tal que $|p - q|$, es decir, el valor absoluto de la *diferencia* de las distancias de P y Q a R sea *máximo*. (Supondremos que L no es la mediatriz de PQ, pues entonces, $p - q$ sería igual a cero para cualquier punto

R de L y el problema carecería de interés.) Para resolverlo, hallemos primero el simétrico de P respecto a L, obteniendo el punto P', del mismo lado de L que Q. Para cualquier punto R' de L, se tiene: $p = R'P = R'P'$, $q = R'Q$. Puesto que R', Q y P' pueden considerarse como los vértices de un triángulo, $|p - q| = |R'P' - R'Q|$ nunca será mayor que $P'Q$, pues la diferencia entre dos lados de un triángulo nunca excede al tercer lado. Si R', P' y Q están sobre la misma recta, $|p - q|$ será *igual* a $P'Q$, como se deduce de la figura. En consecuencia, el punto R buscado es la intersección de L con la recta que une P con Q. Como en el caso anterior, es fácil ver que los ángulos que forman RP y RQ con L son iguales, ya que los triángulos RPR' y y $RP'R'$ son congruentes.

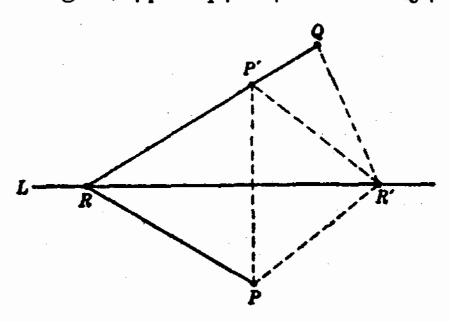


FIG. 182.—$|PR - QR|$ = máximo.

Este problema está relacionado con una propiedad de la tangente a la hipérbola, en igual forma que el anterior lo estaba con la tangente a la elipse. Si la diferencia máxima $|PR - QR|$ tiene el valor $2a$, cabe considerar el lugar geométrico de todos los puntos del plano para los cuales $p - q$ tiene el valor absoluto $2a$; es decir, la hipérbola de focos P y Q que pasa por el punto R. Es fácil ver que el valor absoluto de $p - q$ es menor que $2a$ en la región situada entre las dos ramas de la hipérbola y mayor que $2a$ en el lado de cada rama donde está situado el foco correspondiente. Por un razonamiento esencialmente idéntico al utilizado para la elipse, se deduce que L debe ser tangente a la hipérbola en R. De la situación de P o Q, esto es, de su mayor o menor distancia a L, depende que esta recta sea tangente a una u otra rama de la hipérbola; si P está más cerca, la rama que rodea P tocará a L, y lo mismo para Q (Fig. 183). Si P y Q equidistan de L, entonces ésta no tocará a ninguna de las dos ramas de la hi-

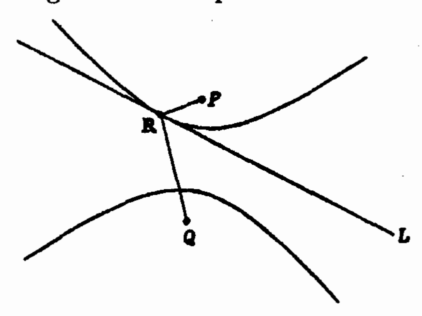


FIG. 183.—Propiedad de la tangente a la hipérbola.

pérbola, sino que será una de las asíntotas de la curva. Este resultado es evidente si se observa que en tal caso la construcción precedente no proporcionará ningún punto R (a distancia finita), ya que la recta $P'Q$ es paralela a L.

Como en los casos anteriores, este razonamiento prueba el teorema bien conocido: la tangente a una hipérbola en cualquier punto de ella es bisectriz del ángulo subtendido en dicho punto por los focos.

Podrá parecer extraño que tengamos que resolver un problema de mínimo cuando P y Q se encuentran del mismo lado de L, mientras que, si están a distinto lado, se transforma en un problema de máximo. Esto es natural, por la siguiente causa: en el primer problema, cada una de las distancias p y q, y, por tanto, su suma, llega a exceder a cualquier límite prefijado, a medida que nos movemos a lo largo de L, en cada una de las dos direcciones posibles. De ahí que sea imposible encontrar un valor máximo para $p + q$, siendo un problema de *mínimo* la única posibilidad. Es enteramente distinto lo que ocurre en el segundo caso, cuando P y Q se encuentran a distinto lado de L. Aquí, para evitar confusiones, debemos distinguir entre el valor de la diferencia $p - q$, su opuesto $q - p$, y el valor absoluto $|p - q|$, siendo este último el que puede ser *máximo*. Se entenderá mejor esto si hacemos que el punto R se mueva a lo largo de L, adoptando distintas posiciones R_1, R_2, R_3, ... Existe un punto para el cual la diferencia $p - q$ es cero: la intersección de la mediatriz de PQ con L. Por consiguiente, este punto corresponde a un mínimo del valor absoluto $|p - q|$. Pero, a un lado de este punto p es mayor que q, y en el otro, menor, de donde resulta que $p - q$ es positiva a un lado de ese punto y negativa al otro. En consecuencia, la propia diferencia $p - q$ no es ni máxima ni mínima en el punto para el cual $|p - q| = 0$. Sin embargo, el punto que hace máximo a $|p - q|$ es realmente un valor extremo de $p - q$. Si $p > q$, se tendrá un máximo para $p - q$; si $q > p$ resultará un máximo para $q - p$; o sea, un mínimo para $p - q$. El que exista máximo o mínimo para $p - q$ queda determinado por la posición de los dos puntos dados, P y Q, respecto a la recta L.

Hemos visto que no existe solución para el problema de máximo si los puntos P y Q equidistan de L, pues entonces, como se observa en la figura 182, $P'Q$ es paralela a L. Esto corresponde al hecho de que la cantidad $|p - q|$ tiende a un límite, cuando R tiende a infinito a lo largo de L, en cualquiera de las dos direcciones posibles. Este valor límite es la longitud de la proyección ortogonal s de PQ sobre L (el lector puede demostrarlo como ejercicio). Si P y Q equi-

distan de L, entonces $|p - q|$ será siempre menor que ese límite, no pudiendo existir máximo, puesto que, para todo punto R, se puede encontrar otro más alejado y tal que el valor correspondiente de $|p - q|$ sea mayor, pero todavía inferior a s.

***5. Distancias extremales a una curva dada.**—Determinaremos, en primer lugar, las distancias *mínima* y *máxima* de un punto P a una curva dada C. Para mayor sencillez, supongamos que C es una curva simple cerrada (Fig. 184) con tangente en cada punto. (Aceptamos aquí el concepto intuitivo de tangente a una curva, el cual será analizado en el próximo capítulo.) La respuesta es muy sencilla: un punto R, situado en la curva C, para el cual la distancia PR sea máxima o mínima, debe ser tal que la recta PR sea perpendicular a la tangente a C en R; en otras palabras, PR debe ser perpendicular a C. La demostración es la siguiente: la circunferencia de centro P, que pasa por R, debe ser tangente a la curva. Pues si R es el punto de distancia mínima, C ha de encontrarse por completo fuera del círculo y, en consecuencia, no puede cortar a su circunferencia en R; por otra parte, si R es el punto de distancia máxima, C debe ser interior a la circunferencia, por lo que tampoco en este caso puede cortarla en R. (Esto se deduce de un hecho evidente: la distancia de cualquier punto a P es menor que RP si el punto es interior al círculo, y mayor que RP si es exterior.) De todo ello resulta que la circunferencia y la curva C se tocarán y tendrán una tangente común en R. Ahora bien: por ser la recta PR un radio de la circunferencia, será perpendicular a su tangente en R y, en consecuencia, perpendicular a C en R.

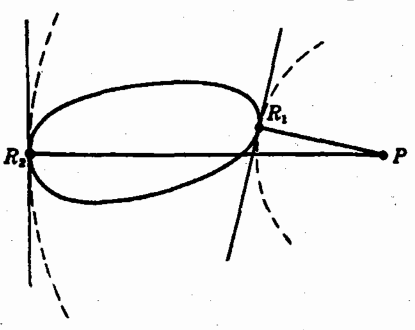


FIG. 184.—Distancias extremales a una curva.

Resulta, además, que el diámetro de una curva cerrada C, es decir, su cuerda máxima, debe ser perpendicular a C en ambos extremos. La demostración queda a cargo del lector, como ejercicio. Una afirmación similar puede formularse y demostrarse para tres dimensiones.

Ejercicio: Demuéstrese que los segmentos máximo y mínimo, que unen dos curvas cerradas que no se cortan, son perpendiculares a las mismas en sus puntos extremos.

Pueden generalizarse ahora los problemas del número anterior, referentes a la suma o diferencia de distancias. En lugar de una recta L

consideremos una curva simple cerrada C, con tangente en cada punto, y dos puntos, P y Q, no situados sobre ella. Deseamos caracterizar los puntos de C, para los cuales la suma $p + q$ y la diferencia $p - q$ toman valores extremos, designando por p y q las distancias de un punto cualquiera de C a P y Q, respectivamente. No es posible utilizar aquí las sencillas construcciones de simetría de que nos servimos cuando C era una recta, pero podemos utilizar las propiedades de la elipse y de la hipérbola para resolver estos problemas. Puesto que C es una curva cerrada y no una recta indefinida, ahora tiene sentido plantearse ambos problemas, el de máximo y el de mínimo, pues puede suponerse que $p + q$ y $p - q$ toman dos valores, máximo y mínimo, sobre cualquier arco finito de curva, en particular sobre una curva cerrada (véanse págs. 376 y siguientes).

En el caso de la suma, $p + q$, supongamos que R es el punto de C para el cual $p + q$ es máximo y sea $2a$ el valor de $p + q$ en R. Consideremos la elipse de focos P y Q, lugar geométrico de todos los puntos tales que $p + q = 2a$. Esta elipse ha de ser tangente a C en R (la demostración queda a cargo del lector, como ejercicio). Pero hemos visto ya que las rectas PR y QR forman ángulos iguales con la elipse en R, y como ésta es tangente a C en R, las rectas PR y QR deben formar también ángulos iguales con C en R. Si $p + q$ es mínimo para R, se ve, de forma análoga, que PR y QR forman ángulos iguales con C en R. Tenemos así este teorema: Se da una curva cerrada C y dos puntos P y Q situados del mismo lado de C; en un punto R de C, en el cual la suma $p + q$ toma su valor máximo o mínimo, las rectas PR y QR forman ángulos iguales con la curva C (es decir, con su tangente) en R.

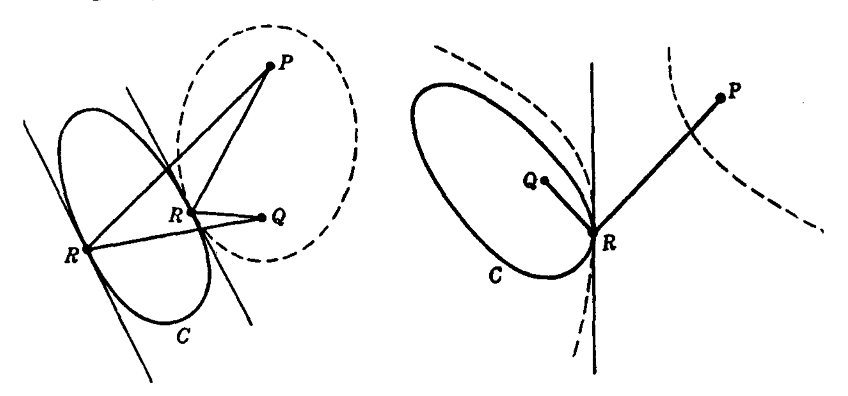


FIG. 185.—Valores máximo y mínimo de $PR + QR$.

FIG. 186.—Mínimo de $PR - QR$.

Si P es interior a C, y Q exterior, el teorema sigue subsistiendo para el valor máximo de $p + q$, pero ya no se cumple para el mínimo, puesto que la elipse degenera en un segmento de recta.

Por un proceso análogo, utilizando las propiedades de la hipérbola en lugar de las de la elipse, el lector podrá demostrar el siguiente teorema: Dada una curva cerrada C y dos puntos P y Q, separados por C, existe un punto R de C en el cual $p - q$ toma un valor máximo o mínimo, y las rectas correspondientes PR y QR forman ángulos iguales con C. Debemos insistir una vez más en que el problema es distinto para una curva cerrada que para una línea indefinida, puesto que en este último caso se buscaba el máximo del valor absoluto $| p - q |$, mientras que ahora existe un máximo (y también un mínimo) de $p - q$.

*II. UN PRINCIPIO GENERAL ACERCA DE LOS PROBLEMAS DE VALORES EXTREMOS

1. El principio.—Los problemas precedentes son ejemplos de una cuestión general, para cuya mejor formulación conviene utilizar el lenguaje analítico. En el problema de hallar los valores extremos de $p + q$, representemos por x e y las coordenadas del punto R; mediante x_1, y_1, las del punto fijo P, y por x_2, y_2, las de Q, con lo que

$$p = \sqrt{(x - x_1)^2 + (y - y_1)^2}, \qquad q = \sqrt{(x - x_2)^2 + (y - y_2)^2},$$

y el problema queda reducido al de determinar los valores extremos de la función

$$f(x, y) = p + q.$$

Esta función es continua en todo el plano, pero el punto cuyas coordenadas son x, y ha de estar situado sobre la curva C, que vendrá definida por una ecuación $g(x, y) = 0$; p. ej., $x^2 + y^2 - 1 = 0$, si se trata de la circunferencia unidad. El problema consiste ahora en encontrar los valores extremos de $f(x, y)$, cuando x e y están sujetos además a la condición $g(x, y) = 0$. En lo que sigue, estudiaremos este tipo general de problema.

Para caracterizar las soluciones, consideremos la familia de curvas definidas por la ecuación $f(x, y) = c$, donde c puede tomar cualquier valor, el mismo para todos los puntos de cada curva de la familia. Supongamos que por cada punto del plano pasa una, y sólo una, de las curvas de la familia $f(x, y) = c$; al menos, si nos limitamos a los puntos situados en la proximidad de la curva C. Entonces, al

variar c, la curva $f(x, y) = c$ barrerá una parte del plano, sin pasar dos veces por un mismo punto de esta región. (Las curvas $x^2 - y^2 = c$, $x + y = c$ y $x = c$ son familias de ese tipo.) En particular, una curva de la familia pasará por R_1, donde $f(x, y)$ toma su valor máximo en C, y otra pasará por el punto R_2, donde $f(x, y)$ toma su valor mínimo. Llamemos a al máximo y b al mínimo. A un lado de la curva $f(x, y) = a$, el valor de $f(x, y)$ será menor que a, y del otro lado, mayor que a. Puesto que $f(x, y) \leqslant a$ en C, ésta debe encontrarse enteramente a un lado de la curva $f(x, y) = a$, por lo que debe ser tangente a ella en R_1. Análogamente, C debe ser tangente a la curva $f(x, y) = b$ en R_2. Tenemos así el teorema general: *si en un punto* R *de una curva* C, *una función* f(x, y) *tiene un valor extremo* a, *la curva* f(x, y) = a *es tangente a* C *en* R.

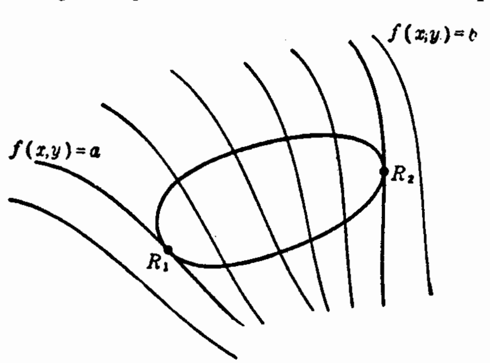


FIG. 187.—Valores extremos de una función sobre una curva.

2. Ejemplos.—Es fácil ver que los resultados anteriores son casos especiales de este teorema general. Si $p + q$ ha de tomar un valor extremo, la función $f(x, y)$ es $p + q$, y las curvas $f(x, y) = c$ son elipses homofocales, de focos P y Q. Como prevé el teorema general, las elipses que pasan por los puntos de C, donde $f(x, y)$ toma sus valores

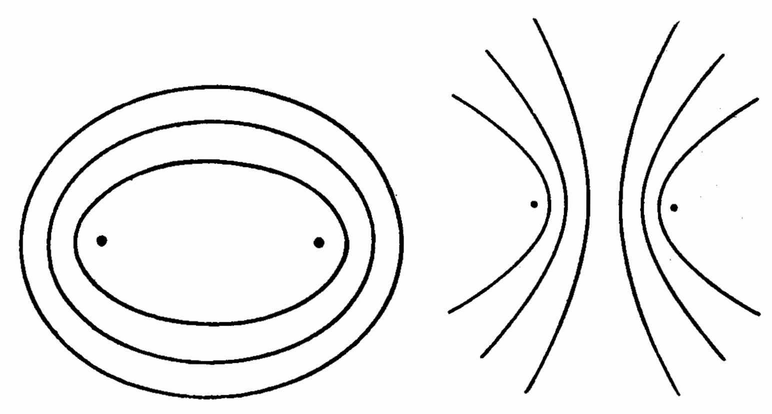


FIG. 188.—Elipses homofocales. FIG. 189.—Hipérbolas homofocales.

extremos, serán tangentes a C en dichos puntos. Cuando se buscan los valores extremos de $p - q$, la función $f(x, y)$ es $p - q$, y las curvas $f(x, y) = c$ son hipérbolas homofocales, de focos P y Q, y las hipérbolas que pasan por los puntos que dan valores extremos a $f(x, y)$ son tangentes a C.

Otro ejemplo es el siguiente: dado un segmento PQ y una recta L, que no lo corta, ¿desde qué punto de L subtenderá PQ un ángulo máximo?

La función cuyo máximo se busca es el ángulo θ que forman los

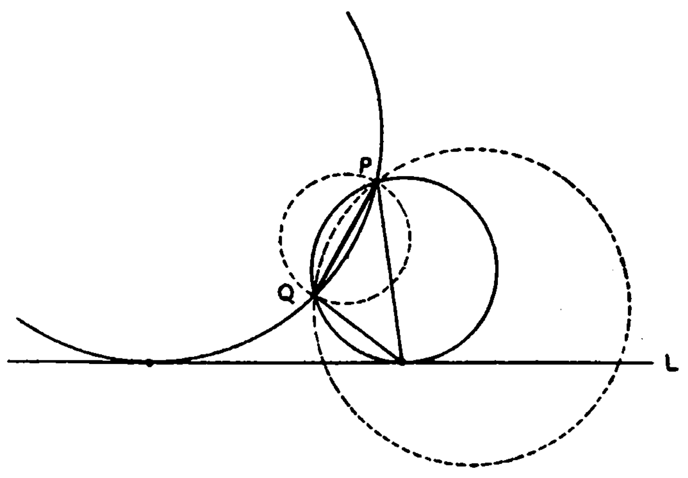


FIG. 190.—Punto de L, desde el que se ve bajo ángulo máximo el segmento PQ.

extremos de PQ con diferentes puntos de L. El ángulo formado por PQ con cualquier punto R del plano es una función $\theta = f(x, y)$ de las coordenadas de R. Por geometría elemental, sabemos que la familia de curvas $\theta = f(x, y) = c$ es el haz de circunferencias que pasan por P y Q, puesto que una cuerda de un círculo subtiende el mismo ángulo para todos los puntos de la circunferencia del mismo lado de la cuerda. Como puede verse en la figura 190, dos de esas circunferencias serán en general tangentes a L, estando sus centros a distinto lado de PQ. Uno de esos puntos de tangencia proporciona el máximo absoluto de θ, mientras que el otro corresponde a un máximo «relativo» (es decir, el valor de θ será, en *un cierto entorno* de este punto, menor que en el propio punto). El mayor de ambos máximos, el máximo absoluto, está dado por el punto de tangencia que se encuentra en el ángulo agudo formado por la prolongación de PQ y L; y el menor, por el punto situado en el ángulo obtuso formado

por estas dos rectas. (El punto donde la prolongación de PQ corta a L proporciona el valor mínimo de θ, o sea, cero.)

Para generalizar este problema, podemos reemplazar L por una curva C y tratar de determinar el punto R de C, tal que un segmento dado PQ (que no corta a C) subtienda un ángulo máximo o mínimo. También en este caso la circunferencia que pasa por P, Q y R debe ser tangente a C en R.

III. LOS PUNTOS ESTACIONARIOS Y EL CÁLCULO DIFERENCIAL

1. Extremos y puntos estacionarios.—Hasta aquí no hemos hecho uso de la técnica del cálculo diferencial. En realidad, nuestros métodos elementales son mucho más sencillos y directos que los del cálculo. Por regla general, en el pensamiento científico, es mejor considerar los rasgos peculiares de un problema que fiarse exclusivamente de los métodos generales, aunque los esfuerzos individuales deban guiarse siempre por un principio que aclare el significado de los procedimientos especiales utilizados. Éste es precisamente el papel que desempeña el cálculo diferencial en los problemas de máximos y mínimos. El ansia moderna por la generalización representa sólo un aspecto de los hechos, pues la vitalidad de la matemática depende primordialmente del matiz individual de los problemas y métodos.

En su desarrollo histórico, el cálculo diferencial fué intensamente influído por los problemas particulares de máximos y mínimos. La conexión entre éstos y el cálculo diferencial se produce de la siguiente manera. En el capítulo VIII haremos un estudio detallado de la derivada $f'(x)$ de una función $f(x)$, y de su significado geométrico. En pocas palabras, la derivada $f'(x)$ es la pendiente de la tangente a la curva $y = f(x)$ en el punto (x, y). Geométricamente, es evidente que en un máximo o mínimo de una curva continua $y = f(x)$, la tangente a la misma debe ser horizontal; es decir, su pendiente igual a cero. Así se obtiene la condición $f'(x) = 0$, que han de cumplir los valores extremos de $f(x)$.

Para comprender lo que significa la anulación de $f'(x)$, examinemos la curva de la figura 191. Existen cinco puntos A, B, C, D y E en los cuales la tangente a la curva es horizontal; designemos por a, b, c, d y e los valores de $f(x)$ en dichos puntos. El máximo de $f(x)$ en el intervalo representado se encuentra en D, y el mínimo, en A. El punto B representa también un máximo en el sentido de que para todos los puntos *en la proximidad inmediata de* B, $f(x)$ es menor que b, aunque $f(x)$ es mayor que b para los puntos que se encuentran cerca

de D. Por esta razón, se dice que B es un *máximo relativo* de $f(x')$ mientras que D es el *máximo absoluto*. Análogamente, C representa un mínimo relativo, y A, el mínimo absoluto. Finalmente, en E no tiene $f(x)$ ni máximo ni mínimo, aunque $f'(x) = 0$. De ahí se deduce que la anulación de $f'(x)$ es una condición *necesaria*, pero no *suficiente*, para que exista un valor extremo de una función derivable

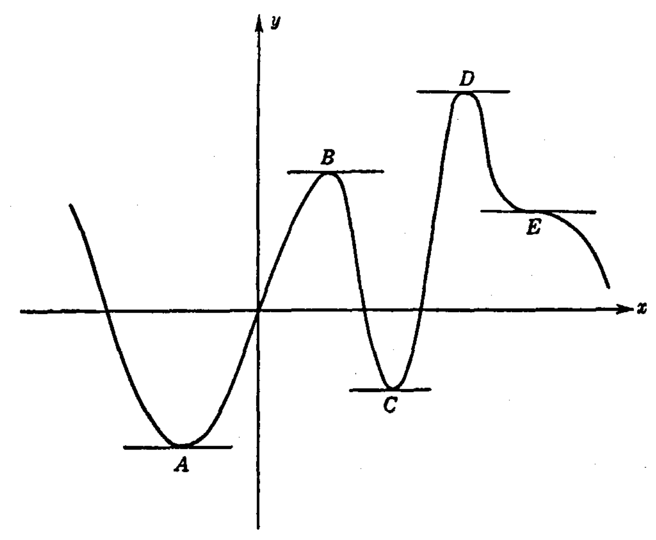


FIG. 191.—Puntos estacionarios de una función.

$f(x)$; en otras palabras, para cualquier valor extremo, relativo o absoluto, debe ser $f'(x) = 0$; pero no en todo punto donde se anula la derivada existe necesariamente un valor extremo de la función. Un punto donde se anula la derivada se llama punto *estacionario*, sea valor extremo o no. Mediante un análisis más profundo, se pueden establecer condiciones más o menos complicadas, en las que se tienen en cuenta las derivadas de orden superior de $f(x)$, que caracterizan por completo los máximos, mínimos y otros puntos estacionarios.

2. Máximos y mínimos de las funciones de varias variables. Puntos de ensilladura.—Existen problemas de máximos y mínimos que no pueden expresarse por medio de una función $f(x)$ de una sola variable. Entre ellos, el más sencillo es el de hallar los extremos de una función $z = f(x, y)$ de dos variables.

Podemos representar $f(x, y)$ por la cota z de una superficie sobre el plano x, y, lo que permite una interpretación, por decirlo así, de geografía de montaña. Un máximo de $f(x, y)$ corresponde a un pico;

un mínimo, al fondo de una depresión o de un lago. En ambos casos, si la superficie no es accidentada, su plano tangente será horizontal. Pero existen otros puntos, además de los picos y de las depresiones, en los cuales el plano tangente es horizontal: son los puntos que representan los pasos de vertiente o puertos, que vamos a examinar con más detalle. Consideremos, como en la figura 192, dos montañas, A y B, que forman parte de una cordillera y dos puntos, C y D, a diferentes lados de la misma; supongamos que se desea ir de C a D. Veamos primero los caminos que conducen de C a D, obtenidos cortando la superficie por un plano que pasa por ambos puntos; cada uno de estos caminos tendrá un punto más alto. Variando la posición del plano cambia el camino, y existirá uno, CD, para el cual la altitud

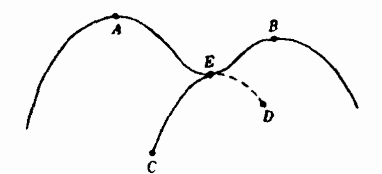


FIG. 192.—Puerto o paso de montaña.

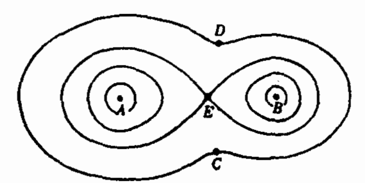


FIG. 193.—Su correspondiente mapa de curvas de nivel.

de su *punto más elevado* es *mínima*. El punto E de máxima altitud de ese camino es un paso de montaña o puerto, que en lenguaje matemático llamaremos *punto de ensilladura*. Es evidente que E no es ni máximo ni mínimo, puesto que tan próximos a él como queramos, podemos encontrar puntos que son más altos o más bajos que E. En lugar de limitarnos a considerar los caminos planos, podemos tener en cuenta otros cualesquiera, prescindiendo de esa restricción. Sin embargo, la caracterización del punto E sigue siendo la misma.

Análogamente, si queremos pasar del pico A al B, cualquier camino que se considere tendrá un punto de mínima altitud; si consideramos otra vez exclusivamente secciones planas, existirá un recorrido AB tal que su punto más bajo sea el más elevado de todos, por lo que, para esta trayectoria, el mínimo se encuentra de nuevo en el punto E anterior. Este punto de ensilladura E tiene la propiedad de ser el mínimo superior o el máximo inferior, es decir, es un *maxi-mínimo* o un *mini-máximo* El plano tangente en E es horizontal; como E es el mínimo de AB, la tangente AB en E debe ser horizontal, y, análogamente, puesto que E es el máximo de CD, la tangente a CD en E ha de serlo también. En consecuencia, el plano tangente, que está determinado por esas rectas, es horizontal. Encontramos así tres tipos

diferentes de puntos, cuyos planos tangentes son horizontales; máximos, mínimos y puntos de ensilladura, a los cuales corresponden los diferentes tipos de valores estacionarios de $f(x, y)$.

Otro método de representar una función $f(x, y)$ consiste en dibujar curvas de nivel, tales como las que se utilizan en los mapas altimétricos (véase pág. 298). Una *curva de nivel* es una curva del plano x, y, a lo largo de la cual la función $f(x, y)$ tiene un valor constante; es decir, las curvas de nivel coinciden con las curvas de la familia $f(x, y) = c$. Por un punto ordinario del plano pasa una curva de nivel y sólo una, mientras que un máximo o mínimo está rodeado por curvas de nivel cada vez más cerradas, y en un punto de ensilladura se cruzan varias de ellas. En la figura 193 se han dibujado las curvas de nivel correspondientes al esquema de la figura 192, siendo evidente la propiedad del punto E de ser maxi-mínimo: cualquier camino que una A con B y que no pase por E ha de atravesar una región en la cual $f(x, y) < f(E)$, mientras que el camino AEB de la figura 192 tiene un mínimo en E. De la misma manera, se ve que el valor de $f(x, y)$ en E es el menor máximo para todos los caminos que unen C con D.

3. Puntos mínimos y topología.—Existe una íntima relación entre la teoría general de los puntos estacionarios y los conceptos topológicos. Sólo podemos dar aquí una indicación muy breve de esas ideas, refiriéndonos a un ejemplo muy sencillo.

Consideremos un paisaje montañoso en una isla (figura 194) en forma de anillo, con sus dos contornos C y C'. Si representamos la altura sobre el nivel del mar por $u = f(x, y)$, con $f(x, y) = 0$ en C y C', y $f(x, y) > 0$ en el interior de la isla, debe existir al menos un paso o puerto en ésta, según aparece en la figura 194, representado por el punto donde se cruzan las líneas de nivel. Intuitivamente es posible ver esto si se intenta ir de C a C' de modo que el camino no pase por ningún punto de altitud mayor que la necesaria. Cada camino de C a C' debe tener un punto de altura máxima, y si elegimos el camino cuyo punto de máxima altitud sea lo más bajo posible, el punto más alto de este recorrido es un punto de ensilladura de $u=f(x, y)$. (Existe una excepción trivial, cuando un plano horizontal es tan-

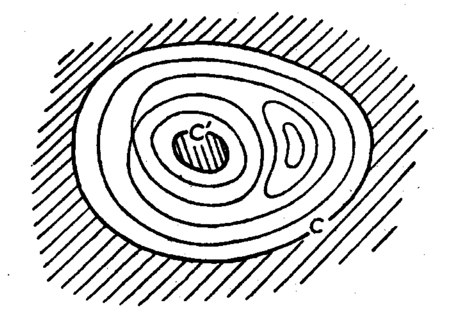


FIG. 194.—Puntos estacionarios en una región doblemente conexa.

gente a la cresta montañosa alrededor de todo el anillo.) Para un dominio limitado por p curvas, deben existir, en general, $p - 1$ puntos estacionarios del tipo estudiado. Marston Morse ha descubierto que esas mismas relaciones son válidas para mayor número de dimensiones, donde existe más variedad de posibilidades topológicas y de tipos de puntos estacionarios. Esas relaciones constituyen la base de la moderna teoría de los puntos estacionarios.

4. Distancia de un punto a una superficie.—Para la distancia entre un punto P y una curva cerrada existen (al menos) dos valores estacionarios: un máximo y un mínimo. No surge ningún hecho nuevo si intentamos generalizar este resultado a tres dimensiones, en tanto que consideremos una superficie C topológicamente idéntica a una

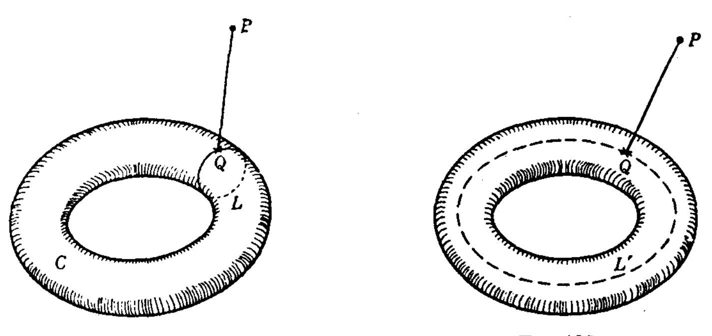


FIG. 195. FIG. 196.

esfera; p. ej., un elipsoide. Pero se presenta una situación distinta si la superficie es de género superior; p. ej., un toro. Existe siempre una distancia máxima y otra mínima, de P al toro C, siendo ambos segmentos perpendiculares a C; pero, además, encontramos valores extremos de diferentes tipos, que son mínimos de máximos o máximos de mínimos. Para ello tracemos en el toro una «meridiana» cerrada, la circunferencia L (Fig. 195), tratando de encontrar el punto Q de L que se encuentre más próximo a P. Después, trasladamos L de modo que la distancia PQ sea: *a)* un mínimo (este Q es simplemente el punto de C que se encuentra más cerca de P); *b)* un máximo (éste nos proporciona otro punto estacionario). También podemos determinar el punto de L que se encuentra más alejado de P, determinando después L de modo que esta distancia máxima sea: *c)* un máximo (que se alcanzará en el punto C más alejado de P); *d)* un mínimo. Obtenemos así cuatro valores estacionarios diferentes de la distancia.

Ejercicio: Repítase el razonamiento anterior para otro tipo L' de curva cerrada sobre C que no pueda reducirse a un punto, como en la figura 196.

IV. EL PROBLEMA DEL TRIÁNGULO DE SCHWARZ

1. La demostración de Schwarz.—Hermann Amandus Schwarz (1843-1921) fué un notable matemático de la Universidad de Berlín y uno de los que más han contribuído a la moderna teoría de funciones y al análisis. No desdeñaba ocuparse de asuntos elementales y una de sus memorias trata del siguiente problema: dado un triángulo acutángulo, inscribirle otro de perímetro mínimo. (Entendemos por triángulo inscrito cualquiera cuyos vértices se encuentran respectivamente en cada uno de los lados del primer triángulo.) Veremos que existe exactamente un triángulo que cumple dicho requisito y que sus vértices coinciden con los pies de las alturas del dado. Le llamaremos *triángulo órtico.*

Schwarz demostró esta propiedad del triángulo órtico de tener el perímetro mínimo mediante el método de reflexión, sirviéndose además del siguiente teorema de geometría elemental (Fig. 197): en cada uno de los vértices P, Q, R los dos lados correspondientes del triángulo órtico forman ángulos iguales con los del primitivo; estos ángulos son iguales al del vértice opuesto del triángulo original; p. ej., los ángulos ARQ y BRP son ambos iguales al C, etc.

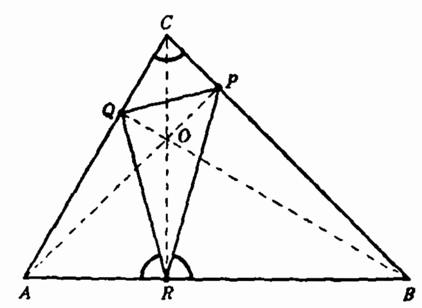


FIG. 197.—Triángulo órtico del ABC; se indican los ángulos que son iguales.

Para demostrar este teorema preliminar, observemos que $OPBR$ es un cuadrilátero inscriptible, puesto que OPB y ORB son ángulos rectos. En consecuencia, $\widehat{PBO} = \widehat{PRO}$, ya que subtienden el mismo arco $\widehat{PO}$ en la circunferencia circunscrita; pero $\widehat{PBO}$ es complementario de $\hat{C}$, por ser CBQ un triángulo rectángulo, y $\widehat{PRO}$ es complementario de $\widehat{PRB}$. En consecuencia, este último es igual a $\hat{C}$. De la misma manera, utilizando el cuadrilátero $QORA$, veríamos que $\widehat{QRA} = \hat{C}$, etc.

Este resultado nos permite enunciar la siguiente propiedad del triángulo órtico: dado que, p. ej., $\widehat{AQR} = \widehat{CQP}$, la simétrica de RQ respecto a AC es la prolongación de PQ, y viceversa; lo mismo sucede para los otros lados.

Demostraremos ahora la propiedad de mínimo del triángulo órtico. En el triángulo ABC consideremos, además del órtico, cualquier otro triángulo inscrito, UVW. Reflejemos toda la figura primero sobre el lado AC de ABC; después repitamos la operación con el triángulo resultante sobre su lado AB, después sobre BC, de nuevo, sobre AC y,

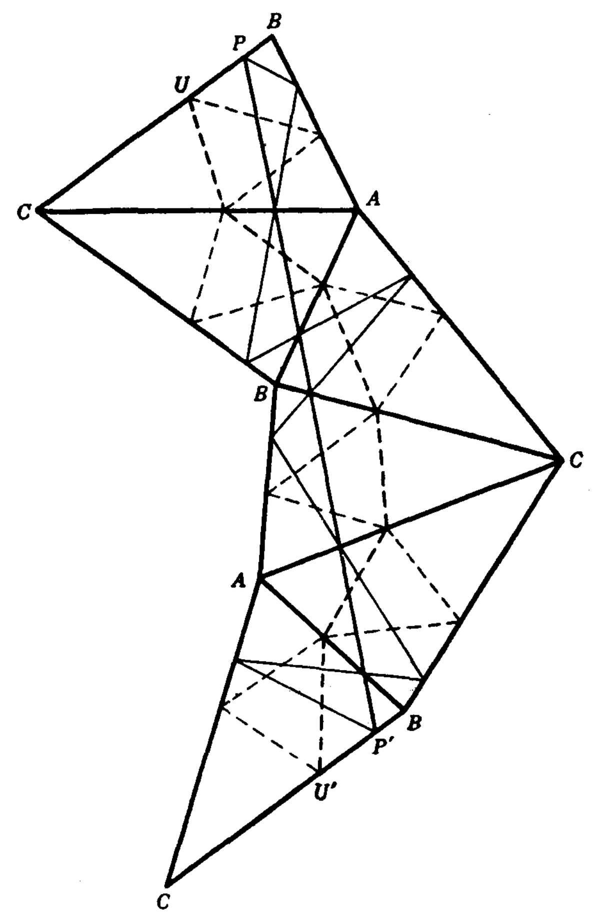


FIG. 198.—Demostración de Schwarz de que el triángulo órtico es el de perímetro mínimo.

finalmente, sobre AB. De esta manera obtenemos en total seis triángulos congruentes. El lado BC del último triángulo es paralelo al primitivo lado BC, pues, en la primera simetría, BC gira en el sentido de las agujas de un reloj un ángulo $2C$; después, en el mismo sen-

tido, un ángulo $2B$; en la tercera simetría no experimenta ningún cambio; en la cuarta, gira un ángulo $2C$ en sentido contrario al de las agujas de un reloj, y en la quinta, un ángulo $2B$ en igual sentido, de donde resulta que el ángulo total girado es cero.

Debido a esta propiedad del triángulo órtico, el segmento PP' es igual a dos veces el perímetro de dicho triángulo; pues PP' se compone de seis trozos que son, respectivamente, iguales al primero, segundo y tercer lados del triángulo, apareciendo cada uno dos veces. Análogamente, la línea quebrada de U a U' es igual al doble del perímetro del otro triángulo inscrito y no es más corta que el segmento UU'. Puesto que éste es paralelo a PP', la quebrada de U

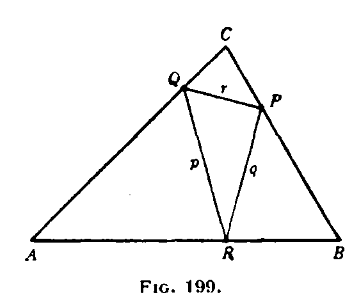


FIG. 199.

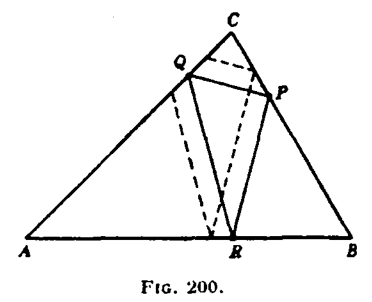


FIG. 200.

a U' no es más corta que PP'; luego el perímetro del triángulo órtico es el menor posible de todos los triángulos inscritos, según se pretendía demostrar. Así hemos deducido al mismo tiempo que existe un mínimo y que está dado por el triángulo órtico. Después demostraremos que no existe ningún otro triángulo de perímetro igual al órtico.

2. **Otra demostración.**—Tal vez la solución más sencilla del teorema de Schwarz sea la que se da a continuación, basada en el teorema antes demostrado de que la suma de las distancias de dos puntos dados, P y Q, a una recta L, es mínima para un punto R de L tal que PR y QR forman ángulos iguales con L, siempre que P y Q se hallen al mismo lado de L y ninguno de ellos esté sobre la recta. Supongamos que el triángulo PQR, inscrito en el ABC, es la solución del problema de mínimo enunciado. Entonces R debe ser un punto del lado AB tal que $p + q$ sea mínimo y, por tanto, los ángulos ARQ y BRP deben ser iguales. Análogamente, se tendrá: $\widehat{AQR} = \widehat{CQP}$, $\widehat{BPR} = \widehat{CPQ}$. Así, pues, si existe el triángulo de perímetro mínimo, ha de verificarse además la igualdad de ángulos utilizada en la demostración de Schwarz. Queda por probar que el único triángulo que

posee tal propiedad es el órtico. Además, puesto que en el teorema en que se basa esta demostración se supone que P y Q no están sobre la recta AB, el razonamiento deja de ser válido cuando uno de los puntos P, Q o R es un vértice del triángulo primitivo (en cuyo caso, el triángulo de perímetro mínimo degeneraría en el doble de la altura correspondiente). Para completar la demostración debemos probar que el perímetro del triángulo órtico es menor que el doble de cualquiera de las alturas.

Para resolver el primer punto observemos que si un triángulo inscrito tiene la propiedad antes mencionada, los ángulos en los vértices P, Q y R deben ser iguales a $\widehat{A}$, $\widehat{B}$, $\widehat{C}$, respectivamente. Pues si suponemos, p. ej., $\widehat{ARQ} = \widehat{C} + \delta$, como la suma de los ángulos de un triángulo es 180°, el ángulo en Q debe ser igual a $B - \delta$ y el ángulo en P, igual a $A - \delta$, para que los ángulos de los triángulos ARQ y BRP den sumas iguales a 180°. Pero la suma de los ángulos del triángulo CPQ es igual a $A - \delta + B - \delta + C = 180° - 2\delta$, y como esta suma ha de ser igual a 180°, necesariamente debe ser $\delta = 0$. Hemos visto ya que el triángulo órtico tiene esta propiedad de la igualdad de ángulos. Cualquier otro triángulo que posea la misma propiedad habrá de tener sus lados paralelos a los correspondientes del triángulo órtico; en otras palabras, sería semejante a él y estaría orientado de la misma manera. El lector demostrará que en un triángulo dado no puede inscribirse otro triángulo cumpliendo tal requisito (Fig. 200).

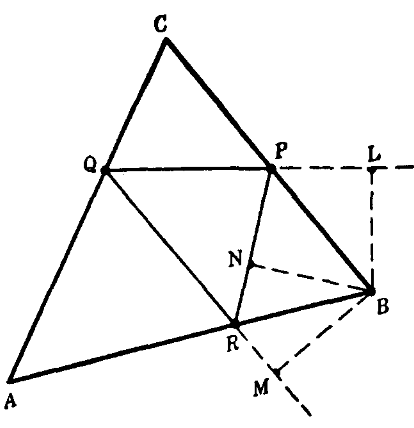


FIG. 201.

Finalmente, demostraremos que el perímetro del triángulo órtico es menor que el doble de cualquiera de las alturas, siempre que los ángulos del triángulo primitivo sean todos agudos. Prolonguemos los lados QP y QR y tracemos las perpendiculares desde B a QP, QR y PR, obteniendo así los puntos L, M y N. Entonces, QL y QM serán, respectivamente, las proyecciones de la altura QB sobre QP y QR. En consecuencia, $QL + QM < 2QB$. Pero $QL + QM$ es igual a p,

perímetro del triángulo órtico, pues los triángulos MRB y NRB son congruentes, por tener iguales los ángulos MRB y NRB y ser rectos los ángulos M y N. De aquí se deduce que $RM = RN$; en consecuencia, $QM = QR + RN$. De la misma manera vemos que $PN = PL$, por lo que $QL = QP + PN$. Tenemos, por tanto, $QL + QM = QP + QR + PN + NR = QP + QR + PR = p$. Pero como hemos demostrado ya que $2QB > QL + QM$, resulta que p es menor que el doble de la altura QB. Con el mismo razonamiento puede demostrarse que p es menor que el doble de cualquier otra altura, con lo cual queda completamente establecida la propiedad de mínimo del triángulo órtico.

Además, la construcción precedente permite calcular directamente p. Sabemos que los ángulos PQC y RQA son iguales a B, por lo que $\widehat{PQB} = \widehat{RQB} = 90° - \widehat{B}$, o sea: $\cos(\widehat{PQB}) = \text{sen}\ \widehat{B}$. En consecuencia, por trigonometría elemental, se deduce que $QM = QL = QB$ sen B y $p = 2QB$ sen B. De igual forma se demuestra que $p = 2PA$ sen $A = 2RC$ sen C. También por trigonometría sabemos que $RC = a$ sen $B = b$ sen A, etc., de donde resulta: $p = 2a$ sen B sen $C = 2b$ sen C sen $A = 2c$ sen A sen B. Finalmente, puesto que $a = 2r$ sen A, $b = 2r$ sen B, $c = 2r$ sen C, siendo r el radio del círculo circunscrito, obtenemos la expresión simétrica $p = 4r$ sen A sen B sen C.

3. Triángulos obtusángulos.—En las dos demostraciones anteriores, se ha supuesto que los ángulos A, B y C eran agudos. Si, p. ej., C es obtuso (Fig. 202), los puntos P y Q se encontrarán fuera del triángulo. En consecuencia, no puede ya decirse estrictamente que el triángulo órtico esté inscrito en el triángulo, a menos que entendamos por triángulo inscrito aquel cuyos vértices se encuentren sobre los lados o sus prolongaciones. En todo caso, el perímetro del triángulo órtico ya no es mínimo, puesto que

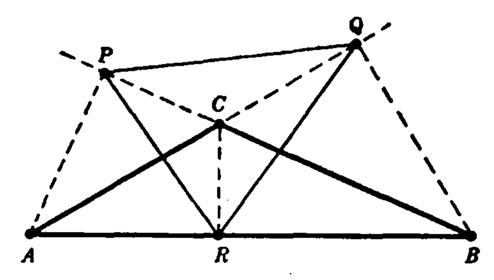


FIG. 202.—Triángulo órtico de un triángulo obtusángulo.

$PR > CR$ y $QR > CR$, de donde se deduce que $p = PR + QR + PQ > 2CR$. Puesto que del razonamiento de la primera parte de la demostración anterior se desprendía que el perímetro mínimo, si no estaba dado por el triángulo órtico, debía ser igual al doble de una de las alturas, concluímos que en los triángulos obtusángulos, el «triángulo inscrito» de perímetro mínimo es la altura menor contada

dos veces, aunque no sea un triángulo propiamente dicho. Además, es posible encontrar un verdadero triángulo cuyo perímetro difiera tan poco como queramos del doble de la altura. Para el caso límite, el triángulo rectángulo, coinciden las dos soluciones: el doble de la altura menor y el triángulo órtico.

No podemos entrar a discutir aquí la interesante cuestión de saber si el triángulo órtico posee propiedades extremales para los triángulos obtusángulos. Nos limitaremos sólo a decir que el triángulo órtico nos proporciona, no un mínimo para la suma de los lados $p+q+r$, sino un valor estacionario de tipo mini-máximo para la expresión $p+q-r$, donde r es el lado del triángulo inscrito opuesto al ángulo obtuso.

4. Triángulos formados por rayos luminosos.—Si el triángulo ABC representa una cámara cuyas paredes pueden reflejar la luz, el triángulo órtico es la única trayectoria luminosa triangular posible. No se excluyen otras trayectorias poligonales, según muestra la figura 203, pero el triángulo órtico es la única poligonal de tres lados.

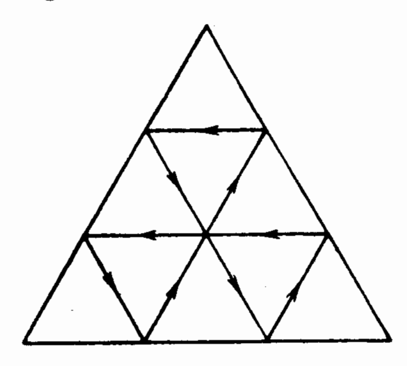
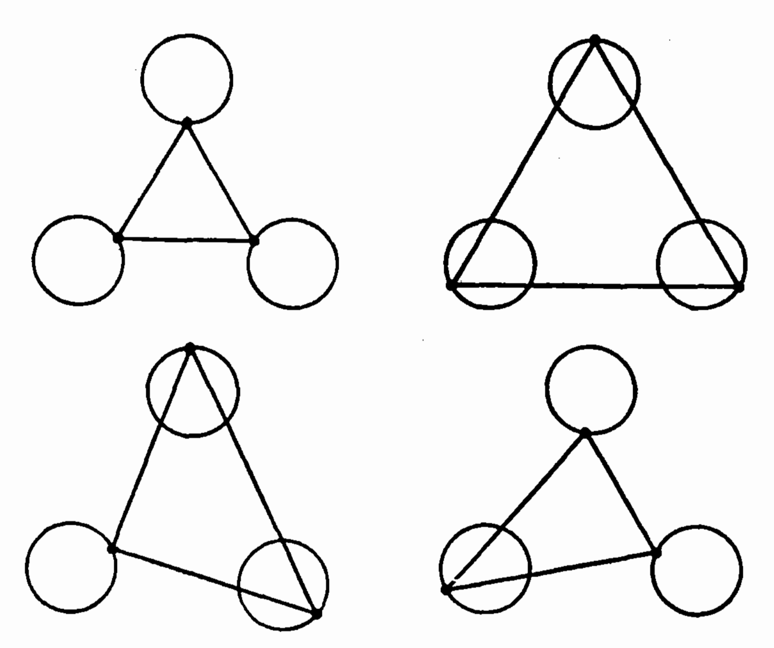


FIG. 203.—Trayectoria luminosa en el interior de un espejo triangular.

Podemos generalizar este problema preguntándonos cuáles son los «triángulos luminosos» posibles en un dominio arbitrario limitado por una o varias curvas continuas; p. ej., deseamos encontrar los triángulos con vértices situados sobre las curvas del contorno y tales que cada dos lados adyacentes formen el mismo ángulo con la curva. Según hemos visto en la página 342, la igualdad de los ángulos es condición necesaria para que exista tanto un máximo como un mínimo de la longitud total de los dos lados, por lo que, de acuerdo con las circunstancias, podremos encontrar diferentes tipos de triángulos luminosos. Así, p. ej., si consideramos el interior de una curva continua única C, el triángulo inscrito de longitud máxima debe ser un triángulo luminoso. O podemos considerar (como nos lo sugiere Marston Morse) el exterior de tres curvas continuas cerradas. Un triángulo luminoso ABC puede caracterizarse por el hecho de que su longitud tenga un valor estacionario; éste puede ser un mínimo respecto a los tres puntos A, B, C, o respecto a cualquiera de las combinaciones, como A y B, y un máximo respecto al tercer punto C, o puede ser un mínimo respecto a un punto y máximo respecto a los otros dos,

o, finalmente, ser un máximo respecto a los tres puntos. En cualquier caso, está asegurada la existencia de, por lo menos, $2^3 = 8$



FIGS. 204-7.—Los cuatro tipos de triángulos luminosos entre tres círculos.

triángulos luminosos, puesto que para cada uno de los tres puntos, independientemente, es posible un máximo o un mínimo.

***5. Observaciones relativas a los problemas de reflexión y al movimiento ergódico.**—Es un problema de especial interés en dinámica y en óptica describir el recorrido o *trayectoria* de una partícula en el espacio, o de un rayo luminoso durante un intervalo ilimitado de tiempo. Si mediante algún artificio físico la partícula o el rayo están restringidos a una porción limitada del espacio, es del mayor interés saber si la trayectoria acabará, en el límite, por llenar todos los puntos del espacio con una distribución aproximadamente uniforme. Una trayectoria de ese tipo se llama *ergódica*. La hipótesis de su existencia es fundamental en los métodos estadísticos de la dinámica moderna y en la teoría atómica. Pero se conocen muy pocos ejemplos pertinentes en los que pueda darse una demostración matemática rigurosa de la *hipótesis ergódica*.

Los ejemplos más sencillos se refieren al movimiento dentro de

una curva plana C, suponiendo que la frontera C se comporta como un espejo perfecto que refleje la partícula (la cual se movería en otro caso sin ninguna restricción) según el mismo ángulo bajo el cual chocó contra la curva; p. ej., una caja rectangular (una mesa de billar ideal de reflexión perfecta y un punto provisto de masa, como una bola) conduce en general a una trayectoria ergódica. La bola ideal, moviéndose indefinidamente, pasará por todo punto, salvo para algunas posiciones iniciales y direcciones singulares. Omitimos la demostración, aunque en principio no es difícil.

De particular interés es el caso de una mesa elíptica, de focos F_1 y F_2. Puesto que la tangente a una elipse forma ángulos iguales con los radios vectores del punto de tangencia, toda trayectoria que pase por un foco se reflejará en el otro, y así indefinidamente. No es difícil ver que, cualquiera que sea la dirección inicial, después de n reflexiones, la trayectoria tiende, al aumentar n, al eje mayor F_1F_2. Si el rayo inicial no pasa por un foco, existen dos posibilidades: si pasa entre ambos focos, todos los rayos reflejados pasarán entre éstos y serán tangentes a una cierta hipérbola, de focos F_1 y F_2; si el rayo inicial no pasa entre F_1 y F_2, tampoco lo hará ninguno de los rayos reflejados y éstos serán tangentes a una elipse de focos F_1 y F_2. Así, pues, el movimiento no será ergódico para la elipse en su totalidad.

***Ejercicios:**

1. Demuéstrese que si el rayo inicial pasa por un foco de la elipse, la n-ésima reflexión del rayo inicial tenderá al eje mayor, al aumentar n.

2. Demuéstrese que si el rayo inicial pasa entre los dos focos, todos los rayos reflejados harán otro tanto y serán tangentes a una hipérbola de focos F_1 y F_2; análogamente, si el rayo inicial no pasa por entre los focos, tampoco lo hará ninguno de los reflejados y todos éstos serán tangentes a una elipse, de focos F_1 y F_2. (Demuéstrese que el rayo, antes y después de reflejarse en R, forma ángulos iguales con las rectas RF_1 y RF_2, respectivamente, y pruébese a continuación que es posible de esta manera caracterizar las tangentes a un sistema de cónicas homofocales.)

V. EL PROBLEMA DE STEINER

1. El problema y su solución.—A principios del siglo XIX, Jacob Steiner, el famoso profesor de Geometría de la Universidad de Berlín, estudió un problema muy sencillo, pero sumamente instructivo. Tres aldeas, A, B, C, han de ser unidas por un sistema de carreteras de longitud total mínima. En términos matemáticos, el problema equivale a que se dan tres puntos A, B, C en el plano y se pide un cuarto punto P tal que la suma $a + b + c$ sea un mínimo, representando esas

letras las distancias de P a A, B y C, respectivamente. La solución del problema es la siguiente: si en el triángulo ABC todos los ángulos son menores de 120°, P es el punto desde el cual cada uno de los tres lados AB, BC, CA subtiende un ángulo de 120°. Sin embargo, si un ángulo de ABC, p. ej., el C, es igual o mayor que 120°, el punto P coincide con el vértice C.

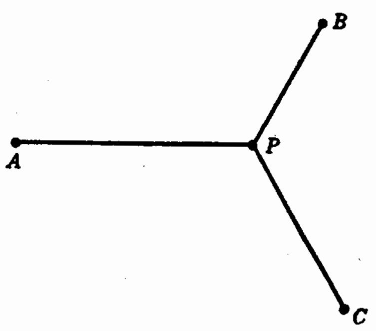


Fig. 208. — Suma mínima de distancias a tres puntos.

Es fácil obtener esta solución, utilizando los resultados anteriores relativos a valores extremos. Supongamos que P es el punto buscado. Existen las siguientes posibilidades: o P coincide con uno de los vértices A, B, C, o es distinto de ellos. En el primer caso, es evidente que P debe ser el vértice del mayor ángulo C de ABC, puesto que la suma $CA + CB$ es menor que cualquier otra suma de otros dos lados del triángulo ABC. Así, para completar la demostración, debemos considerar el segundo caso. Sea K una circunferencia de radio c y centro C. Entonces P debe ser un punto de K, tal que $PA + PB$ sea mínimo. Si A y B son exteriores a K (Fig. 209), de acuerdo con lo dicho en la página 342, PA y PB deben formar ángulos iguales con la circunferencia K, o sea, con el radio PC, que es perpendicular a K. Puesto que el mismo razonamiento se aplica a la posición de P y al círculo de radio a y centro A, resulta que los tres ángulos formados por PA, PB, PC son iguales, o sea, que cada uno vale 120° como habíamos dicho. En el razonamiento se ha supuesto que A y B son exteriores a K, lo que queda aún por demostrar. Si al menos uno de los puntos A y B, p. ej., A, se encontrara en K o fuera interior, puesto que P, como se ha supuesto, no coincide con A o B, tendríamos $a + b \geqslant AB$. Pero $AC \leqslant c$, ya que A no es exterior a K. De donde resulta:

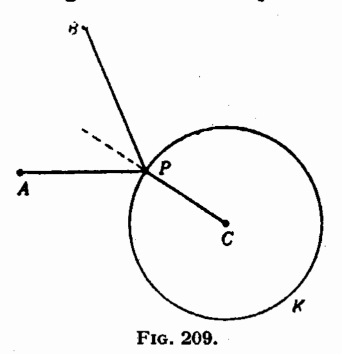


Fig. 209.

$$a + b + c \geqslant AB + AC,$$

lo que significa que obtendríamos la mínima suma de distancias si P coincidiera con A, contra nuestra hipótesis. Esto demuestra que tanto A como B son exteriores al círculo K. El hecho correspondiente se demuestra en forma similar para las otras combinaciones: B, C respecto al círculo de radio a y centro A, y A, C, respecto al círculo de radio b y centro B.

2. Análisis de los casos posibles.—Para verificar cuál de los dos casos posibles respecto a la posición de P sucede realmente debemos examinar la construcción de dicho punto P. Para determinar éste, trazamos las circunferencias K_1 y K_2, sobre las cuales dos lados, p. ej., AC y BC, subtienden arcos de 120°. Entonces, AC subtenderá también un ángulo de 120° desde cualquier punto del menor de los arcos en que AC divide a K_1; pero subtenderá 60° desde cualquier punto

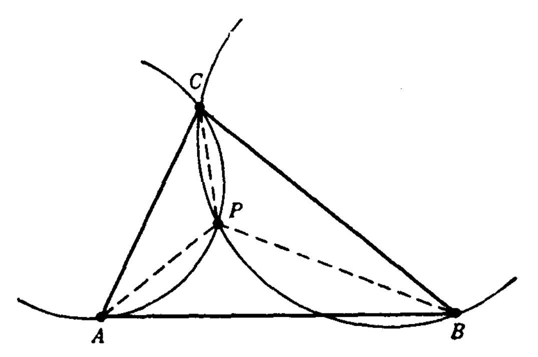


FIG. 210.

del arco mayor. La intersección de los dos arcos menores, siempre que exista, proporciona el punto buscado P, pues no sólo AC y BC subtenderán un arco de 120° en P, sino que lo mismo sucederá con AB, por ser 360° la suma de los tres ángulos.

De la figura 210 resulta evidente que si ninguno de los ángulos del triángulo ABC es mayor que 120°, los dos arcos más cortos se cortarán dentro del triángulo. Por otra parte, si un ángulo C del triángulo ABC es mayor que 120°, los dos arcos más cortos de K_1 y K_2 no se cortan, como se ve en la figura 211. En este caso, no existe un punto P desde el cual los tres lados subtiendan ángulos de 120°. Sin embargo, K_1 y K_2 determinan, por su intersección, un punto P', desde el cual AC y BC subtienden ángulos de 60° cada uno, mientras el lado AB, opuesto al ángulo obtuso, subtiende un ángulo de 120°.

Para un triángulo ABC que tenga un ángulo mayor de 120° no

existe, pues, ningún punto desde el cual los tres lados subtiendan ángulos de 120°. De ahí que el punto mínimo P deba coincidir con un vértice, puesto que, como ya se ha demostrado, es la única posi-

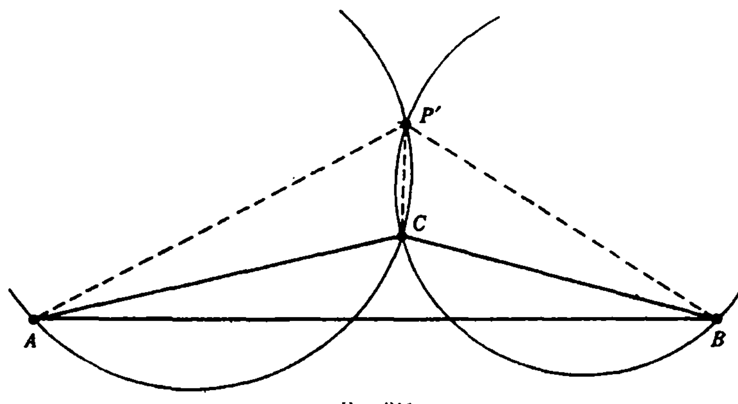


FIG. 211.

bilidad restante, debiendo ser además el vértice del ángulo obtuso. Si, por otra parte, todos los ángulos del triángulo son menores de 120°, hemos visto que se puede determinar un punto desde el cual cada lado subtienda un ángulo de 120°. Pero para completar la demostración debemos probar todavía que $a + b + c$ será realmente menor, en este caso, que si P coincidiera con cualquiera de los vértices, ya que sólo hemos demostrado que P proporciona un mínimo *si* la menor longitud total no corresponde a uno de los vértices. En consecuencia, debemos demostrar que $a + b + c$ es menor que la suma de dos lados cualesquiera, p. ej., $AB + AC$. Para conseguirlo, prolonguemos BP y proyectemos A sobre esta recta, con lo que se obtiene un punto D (Fig. 212). Puesto que $\widehat{APD} = 60°$, la longitud de la proyección PD es $\frac{1}{2}a$, y como BD es la proyección de AB sobre BP, se tendrá que $BD < AB$. Ahora bien: $BD = b + \frac{1}{2}a$; o sea,

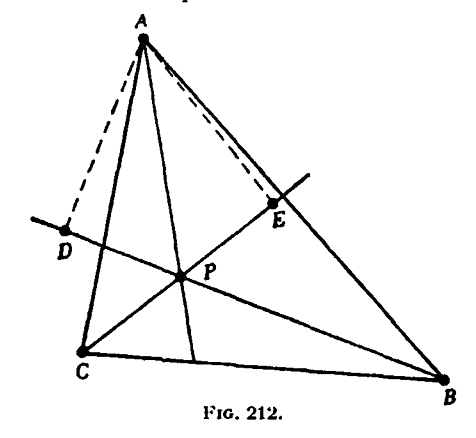


FIG. 212.

$b + {}^1/_2 a < AB$. De igual forma, proyectando A sobre la prolongación de PC, veríamos que $c + {}^1/_2 a < AC$. Sumando, obtenemos la desigualdad $a + b + c < AB + AC$. Puesto que ya sabemos que el punto mínimo, si no es uno de los vértices, debe ser P, se deduce finalmente que éste es realmente el punto para el cual $a + b + c$ es mínima.

3. Un problema complementario.—Los métodos formales de la matemática nos llevan algunas veces más allá del objetivo original; p. ej., si el ángulo C es mayor de 120°, la construcción geométrica determina, no el punto solución P (que en este caso es el propio punto C), sino otro punto P' desde el cual aparece el lado mayor AB del triángulo ABC bajo un ángulo de 120°, y los dos más pequeños se ven bajo ángulos de 60° cada uno. Ciertamente, P' no resuelve nuestro problema de mínimo, pero cabe sospechar que tiene alguna relación con él. La respuesta es que P' resuelve el problema siguiente: hallar el mínimo de la expresión $a+b-c$. La demostración es por completo análoga a la dada para $a + b + c$, basada en los resultados obtenidos en la página 347, por lo que se deja como ejercicio al lector. Si se combina con el resultado precedente, tenemos el teorema:

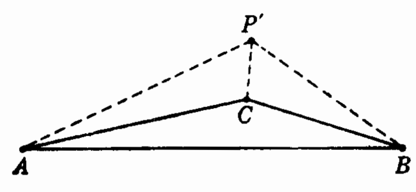


FIG. 213.—$a + b - c =$ mínimo.

Si los ángulos de un triángulo ABC son todos menores que 120°, la suma de las distancias a, b, c de un punto cualquiera a A, B, C, respectivamente, es mínima para aquel punto desde el cual se ve cada lado del triángulo bajo un ángulo de 120°, y $a + b - c$ es mínima para el vértice C; si un ángulo, p. ej., C, es mayor que 120°, $a + b + c$ es mínima para C, y $a + b - c$ lo es para un punto tal que los dos lados menores del triángulo subtiendan ángulos de 60°, mientras que el mayor subtienda otro de 120°.

Así, pues, de los dos problemas de mínimo, uno puede resolverse siempre por la construcción conocida, y la solución del otro es un vértice. Para $\widehat{C} = 120°$, las dos soluciones de cada problema, e incluso las soluciones de ambos, coinciden, puesto que el punto que se obtiene por construcción es precisamente el vértice C.

4. Observaciones y ejercicios.—Si desde un punto P, situado en el interior de un triángulo equilátero UVW, se trazan tres rectas perpendiculares PA, PB, PC, según muestra la figura 214, los puntos

A, B, C y P forman la figura estudiada anteriormente. Esta observación puede servir para resolver el problema de Steiner, comenzando por los puntos A, B, C y determinando a continuación U, V, W.

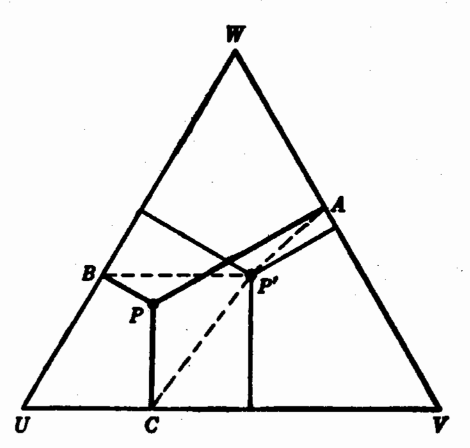


FIG. 214.—Otra demostración de la solución de Steiner.

Ejercicios:

1. Efectúese esta construcción, utilizando la propiedad de que para todo punto de un triángulo equilátero, la suma de los tres segmentos normales a sus lados es constante e igual a la altura.

2. Haciendo uso de la propiedad correspondiente cuando P es exterior al triángulo, discútase el problema complementario. En tres dimensiones puede estudiarse el problema análogo siguiente: dados cuatro puntos A, B, C, D, hállese un quinto punto P tal que $a + b + c + d$ sea mínima.

***Ejercicio:** Estúdiese este problema y su complementario mediante los métodos dados en las páginas 341 y siguientes, o utilizando un tetraedro regular.

5. Generalización al problema de la red de carreteras.—En el problema de Steiner, se dan tres puntos A, B, C. Es natural generalizarlo al caso de ser n los puntos dados en el plano, A_1, A_2, A_3, ... , A_n, y pidiéndose entonces el punto P, para el cual la suma de las distancias $a_1 + a_2 + a_3 + ... + a_n$ sea mínima, siendo a_i la distancia PA_i. (Para cuatro puntos, dispuestos como en la figura 215, el punto P es el de intersección de las diagonales del cuadrilátero $A_1A_2A_3A_4$, lo que el lector puede demostrar como ejercicio.)

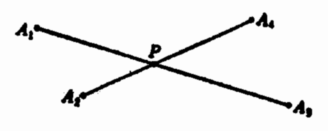
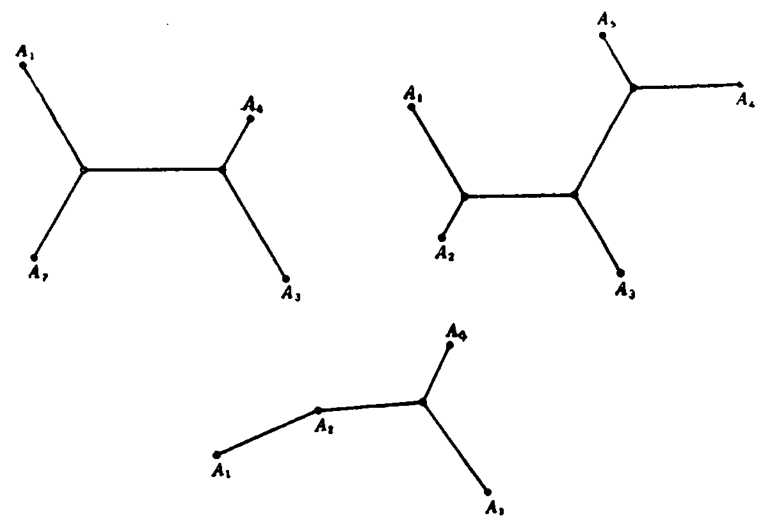


FIG. 215.—Suma mínima de distancias a cuatro puntos.

Este problema, del que también se ocupó Steiner, nos conduce a resultados interesantes. Es una de tantas generalizaciones superficiales que aparecen con frecuencia en la literatura matemática. Para encontrar la generalización verdaderamente significativa del problema de Steiner debemos abandonar la búsqueda de un punto único P, y en su lugar trataremos de hallar la «red de carreteras» de longitud total mínima, lo que en lenguaje matemático se expresa así: *Dados* n *puntos* A_1, ..., A_n *determínese un sistema conexo de segmentos rectilíneos*

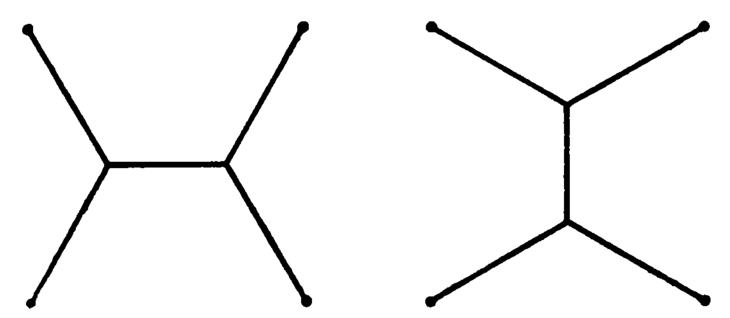
de longitud total mínima, de tal modo que dos puntos cualesquiera queden unidos por una poligonal formada por segmentos del sistema.

La solución depende, por supuesto, de la ordenación de los puntos



FIGS. 216-18. — Redes mínimas uniendo más de tres puntos.

dados. El lector podrá estudiar fructíferamente el tema, basándose en la solución del problema de Steiner. Aquí nos limitaremos a esbozar la respuesta en los casos típicos representados en las figuras 216 a 218. En el primero, la solución consta de cinco segmentos con dos



FIGS. 219-20. — Dos redes mínimas uniendo cuatro puntos.

intersecciones múltiples, donde tres de los segmentos se cortan formando ángulos de 120°. En el segundo caso, la solución contiene tres intersecciones múltiples. Si los puntos se disponen de otra manera, dichas figuras pueden ser imposibles. Una o varias de las intersec-

ciones múltiples pueden degenerar, apareciendo en su lugar uno o más de los puntos dados, como en el tercer caso.

En el caso de ser n los puntos dados, existirán por lo menos $n-2$ intersecciones múltiples, en cada una de las cuales tres segmentos formarán ángulos de 120°.

La solución del problema no está siempre unívocamente determinada. Para cuatro puntos A, B, C, D, que formen un cuadrado, tenemos dos soluciones equivalentes que aparecen en las figuras 219 y 220. Si los puntos $A_1, A_2, \ldots, A_n$ son los vértices de una poligonal simple con ángulos suficientemente obtusos, esta misma será la de longitud mínima.

VI. VALORES EXTREMOS Y DESIGUALDADES

Uno de los rasgos característicos de la matemática superior es el importante papel que las desigualdades desempeñan en ella. En principio, la solución de un problema de máximo conduce siempre a una desigualdad, la cual expresa el hecho de que la variable que se considera es menor o a lo sumo igual al valor máximo que proporciona la solución. En muchos casos, tales desigualdades tienen un interés independiente por sí mismas. Como ejemplo, consideraremos la importante desigualdad entre las medias aritmética y geométrica.

1. Medias aritmética y geométrica de dos cantidades positivas. Comencemos por un sencillo problema de máximo que aparece a menudo en la matemática pura y en la aplicada. En lenguaje geométrico equivale a lo siguiente: entre todos los rectángulos de perímetro dado, hállese el de mayor área. Como era de esperar, la solución es un cuadrado. Para demostrarlo se razona de la manera siguiente: sea $2a$ el perímetro dado del rectángulo. La suma de dos lados adyacentes, $x + y$, queda fijada al dar el perímetro, mientras que el área variable xy deberá hacerse tan grande como sea posible. La «media aritmética» de x e y es sencillamente

$$m = \frac{x + y}{2}.$$

Introduzcamos ahora la cantidad

$$d = \frac{x - y}{2},$$

de donde se deduce:

$$x = m + d, \qquad y = m - d,$$

y, en consecuencia,

$$xy = (m + d)(m - d) = m^2 - d^2 = \frac{(x + y)^2}{4} - d^2.$$

Puesto que d^2 es mayor que cero, excepto para $d = 0$, se obtiene inmediatamente la desigualdad

$$\sqrt{xy} \leqslant \frac{x+y}{2}, \tag{1}$$

donde el signo de igualdad sólo es válido cuando $d=0$ y $x=y=m$.

Puesto que $x + y$ es fija, se deduce que $\sqrt{xy}$, y, en consecuencia, el área xy será máxima, cuando $x = y$. La expresión

$$g = \sqrt{xy},$$

en la que deberá tomarse la raíz con signo positivo, se llama «media geométrica» de las cantidades positivas x e y. La desigualdad [1] expresa la relación fundamental entre las medias aritmética y geométrica.

Se obtiene también directamente la desigualdad [1] del hecho de ser necesariamente no negativa la expresión

$$(\sqrt{x} - \sqrt{y})^2 = x + y - 2\sqrt{xy},$$

pues es un cuadrado que sólo puede ser igual a cero para $x = y$.

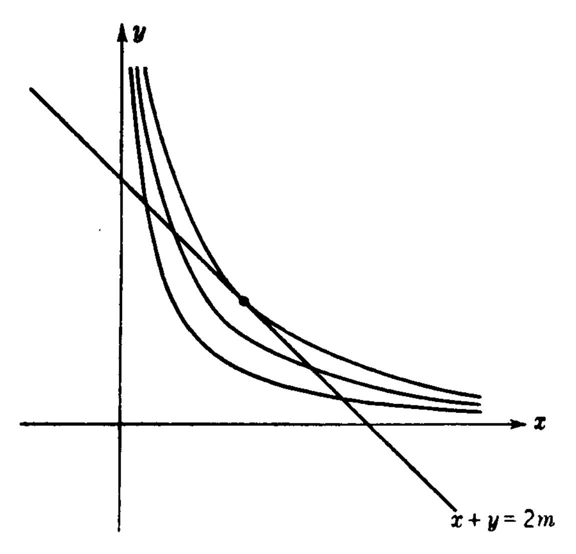


FIG. 221.—Máximo de xy, dada $x + y$.

Puede deducirse una consecuencia geométrica de dicha desigualdad, considerando en el plano la recta fija $x + y = 2m$, junto con la familia de curvas $xy = c$, donde c es constante para cada una de

las curvas (hipérbolas), y varía de una a otra. Como resulta evidente de la figura 221, la curva para la que es máximo el valor de c y tiene un punto común con la recta dada es la hipérbola tangente a la recta en el punto $x = y = m$; para ella, evidentemente, $c = m^2$. En consecuencia,

$$xy \leqslant \left(\frac{x+y}{2}\right)^2$$

Debe hacerse constar que cualquier desigualdad del tipo $f(x, y) \leqslant \leqslant g(x, y)$ puede leerse en ambas direcciones, por lo que proporciona no sólo una propiedad de máximo, sino también otra de mínimo; p. ej., [1] expresa también la propiedad de que entre todos los rectángulos de área dada, el cuadrado es el de perímetro mínimo.

2. Generalización para «n» variables.—Puede generalizarse la desigualdad [1] entre las medias aritmética y geométrica para cualquier número n de cantidades positivas, que denotaremos por $x_1, x_2, x_3, ..., x_n$. Decimos que

$$m = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

es la media aritmética, y

$$g = \sqrt[n]{x_1x_2 \cdots x_n},$$

la media geométrica, entendiendo en este último caso que se trata de la raíz positiva. El teorema general dice que

$$g \leqslant m, \qquad [2]$$

siendo $g = m$ sólo si todos los x_i son iguales.

Se han dado muchas y muy ingeniosas demostraciones de este resultado general. La más sencilla consiste en reducirla al mismo razonamiento utilizado en la página 371, planteando el siguiente problema de máximo: dividir una cantidad positiva C en n partes positivas, $C = x_1 + x_2 + ... + x_n$, de tal modo que el producto $P = = x_1x_2 ... x_n$ sea máximo. Comencemos por suponer—hipótesis en apariencia obvia, pero que se analizará más adelante—que existe un máximo de P y que éste tiene lugar para el siguiente conjunto de valores:

$$x_1 = a_1, \cdots, x_n = a_n.$$

Todo se reduce a probar que $a_1 = a_2 = ... = a_n$, pues en tal caso

se tiene: $g = m$. Supongamos que eso no fuera cierto; p. ej., que $a_1 \neq a_2$. Consideremos las n cantidades

$$x_1 = s, \qquad x_2 = s, \qquad x_3 = a_3, \cdots, x_n = a_n,$$

siendo

$$s = \frac{a_1 + a_2}{2}$$

En otras palabras, reemplacemos las cantidades a_i por otro conjunto, en el cual han cambiado sólo las dos primeras, que se han hecho iguales, mientras que la suma total C permanece invariable. Podemos escribir

$$a_1 = s + d, \qquad a_2 = s - d,$$

donde

$$d = \frac{a_1 - a_2}{2}$$

El nuevo producto es

$$P' = s^2 \cdot a_3 \cdots a_n,$$

mientras que el primitivo era

$$P = (s + d) \cdot (s - d) \cdot a_3 \cdots a_n = (s^2 - d^2) \cdot a_3 \cdots a_n,$$

por lo que, evidentemente, a menos que sea $d = 0$, se tendrá

$$P < P',$$

contra la hipótesis según la cual P era máximo. En consecuencia, $d = 0$ y $a_1 = a_2$. De la misma manera podemos demostrar que $a_1 = a_i$, donde a_i representa cualquiera de las cantidades a. Finalmente, se deduce que todas las a_i deberán ser iguales. Puesto que $g = m$ cuando todas las x_i son iguales, y como hemos demostrado que sólo en este caso se obtiene el máximo de g, resulta que, en cualquier otro, será $g < m$, como queríamos demostrar.

3. El método de los cuadrados mínimos.—La media aritmética de n números $x_1, x_2, ..., x_n$, que en lo que sigue no precisan ser necesariamente positivos, tiene una importante propiedad de mínimo. Sea u una cantidad desconocida, que queremos determinar con tanta precisión como sea posible, mediante un instrumento de medida. Con este fin, efectuamos cierto número n de lecturas, que pueden dar lugar a resultados ligeramente distintos, $x_1, ..., x_n$, debido a diversas causas de error experimental. Surge entonces la cuestión de saber qué valor de u deberá aceptarse como más *plausible;* es costumbre

elegir como valor «plausible» u «óptimo» la media aritmética $m = \frac{x_1 + \cdots + x_n}{n}$. Para justificar debidamente esta hipótesis, sería necesario entrar en una detallada discusión de la teoría de las probabilidades. Pero, al menos, podemos señalar una propiedad de mínimo de m que hace razonable nuestra elección. Sea u un valor cualquiera posible de la cantidad medida. Las diferencias $u - x_1, \ldots, u - x_n$, medirán las desviaciones existentes entre las diferentes lecturas x_i y este valor. Algunas de esas desviaciones serán positivas, mientras otras serán negativas, y parece natural suponer que el valor óptimo de u es tal que la desviación total, en cierto sentido, resulte lo más pequeña posible. Siguiendo a Gauss, se acostumbra tomar, no las desviaciones, sino sus cuadrados $(u - x_i)^2$, como una medida apropiada del error, y elegir como valor óptimo entre todas los posibles valores de u el que hace mínima la suma de los cuadrados de las desviaciones:

$$(u - x_1)^2 + (u - x_2)^2 + \cdots + (u - x_n)^2.$$

Este valor óptimo de u *es precisamente la media aritmética* m, hecho que constituye el punto de partida del importante «método de los cuadrados mínimos» de Gauss. Podemos demostrar de una manera muy elegante la afirmación en cursiva. Si escribimos

$$(u - x_i) = (m - x_i) + (u - m),$$

se obtiene:

$$(u - x_i)^2 = (m - x_i)^2 + (u - m)^2 + 2(m - x_i)(u - m).$$

Sumemos ahora todas estas ecuaciones para $i = 1, 2, \ldots, n$. Los últimos términos dan $2(u - m)(nm - x_1 - \ldots - x_n)$, que es cero, como se deduce de la definición de m. En consecuencia, queda

$$(u - x_1)^2 + \cdots + (u - x_n)^2 = (m - x_1)^2 + \cdots + (m - x_n)^2 + n(m - u)^2,$$

lo que demuestra que

$$(u - x_1)^2 + \cdots + (u - x_n)^2 \geqslant (m - x_1)^2 + \cdots + (m - x_n)^2,$$

y que el signo de igualdad sólo se verifica para $u = m$, según queríamos probar.

El método general de los cuadrados mínimos parte de este resultado, utilizándolo como norma orientadora en los casos más complicados, cuando se trata de decidir cuál es el valor más plausible que puede deducirse de medidas entre las cuales existen ligeras contradicciones. Supongamos, p. ej., que hemos medido las coordenadas de n

puntos, x_i, y_i, de una línea que teóricamente debía ser recta, y supongamos además que los puntos medidos no se encuentran exactamente sobre una recta. ¿Cómo deberemos trazar la recta que mejor se *ajuste* a los n puntos observados? Los resultados anteriores sugieren el procedimiento siguiente, que, ciertamente, podría reemplazarse por otras variantes igualmente razonables: sea $y = ax + b$ la ecuación de la recta, por lo que el problema consiste en determinar los coeficientes a y b. La distancia en la dirección de las y, desde la recta al punto x_i, y_i, está dada por $y_i - (ax_i + b) = y_i - ax_i - b$, siendo el signo positivo o negativo, según que el punto se encuentre por encima o por debajo de la recta. El cuadrado de esa distancia será, pues, $(y_i - ax_i - b)^2$ y el método consiste sencillamente en determinar a y b de tal manera que la expresión

$$(y_1 - ax_1 - b)^2 + \cdots + (y_n - ax_n - b)^2$$

tenga el menor valor posible. Así, pues, en este caso tenemos un problema de mínimo que incluye dos cantidades desconocidas: a y b. Aunque muy sencilla, omitimos la discusión detallada de la solución.

VII. EXISTENCIA DE EXTREMOS. PRINCIPIO DE DIRICHLET

1. Observaciones generales.—En algunos de los problemas anteriores de máximos y mínimos se prueba directamente que la solución proporciona el mejor resultado posible. Un ejemplo notable es la solución de Schwarz del problema del triángulo, en cuyo caso pudimos ver sin dificultad que ningún triángulo inscrito tiene un perímetro menor que el triángulo órtico. Otros ejemplos son los problemas de máximos y mínimos, cuyas soluciones dependen de una desigualdad explícita, tal como la existente entre las medias geométrica y aritmética. Pero en algunos de los problemas tratados hemos seguido un camino distinto. Se comienza por suponer que se ha encontrado una solución; analizando esta hipótesis extraemos conclusiones que pueden eventualmente caracterizar y construir la solución. Así ocurrió, por ejemplo, con la solución del problema de Steiner y con el segundo método de resolución del problema de Schwarz. Ambos métodos son lógicamente diferentes. En cierto sentido, el primero es más perfecto, puesto que proporciona una demostración, más o menos constructiva, de la solución. Como tuvimos ocasión de ver en el caso del problema del triángulo, es probable que el segundo método sea más sencillo; pero no es tan directo y, sobre todo, tiene un vicio de origen en su estructura, pues empieza suponiendo que *existe* una solución del pro-

blema. Proporciona tal solución *sólo* si se admite o se demuestra que ésta existe. Sin esa hipótesis previa, se limita a probar que *si* existe una solución, debe poseer cierto carácter[1].

Debido a la aparente evidencia de la premisa de que existe una solución, los matemáticos, hasta finales del siglo XIX, no prestaron atención a la cuestión lógica inmanente, aceptando como cosa natural la existencia de solución en los problemas de máximos y mínimos. Algunos de los más grandes matemáticos del siglo XIX—Gauss, Dirichlet y Riemann—utilizaron sin discriminación tal hipótesis, tomándola como base para la investigación de difíciles teoremas de física matemática y teoría de funciones, que de otro modo hubieran resultado casi inasequibles. El punto culminante se produjo en 1849, cuando Riemann publicó su tesis doctoral sobre los fundamentos de la teoría de funciones de variable compleja. Aquella memoria, tan concisamente escrita y uno de los trabajos más importantes de la matemática moderna, era tan heterodoxa en su forma de atacar el problema, que mucha gente hubiera preferido ignorarla. Weierstrass era entonces el matemático más notable de la Universidad de Berlín, y el maestro reconocido por todos de la construcción de una teoría de funciones rigurosa. Impresionado por su lectura, pero un tanto escéptico, pronto descubrió una laguna lógica en aquella memoria, que el autor no se había preocupado de llenar. La rigurosa crítica de Weierstrass, si bien no influyó en el ánimo de Riemann, hizo que se ignorara casi por completo su teoría. La carrera meteórica de Riemann terminó súbitamente, pues murió tuberculoso pocos años después. Pero sus ideas encontraron siempre algunos entusiastas discípulos, y cincuenta años después de la publicación de su tesis, Hilbert consiguió abrir el camino para dar una respuesta completa a todas las cuestiones que había dejado sin esclarecer. Todo este desarrollo de la matemática y de la física matemática se convirtió en uno de los más grandes triunfos en la historia del análisis matemático moderno.

En la memoria de Riemann, el punto expuesto al ataque de la crítica es la cuestión de la existencia de mínimo. Riemann basó gran parte de su teoría en lo que él llamaba *principio de Dirichlet.* (Dirichlet había sido profesor de Riemann en Gotinga, y había expuesto esta cuestión en sus clases, pero nunca publicó nada sobre dicho principio.) Supongamos, p. ej., que una parte de un plano o de cualquier superficie se recubre con una lámina de estaño muy fina, y que se establece una corriente eléctrica estacionaria en la lámina metálica, co-

[1] La necesidad lógica de probar la existencia de un extremo se aclara con el siguiente sofisma: 1 es el mayor de los enteros. Representemos por x el máximo entero; si fuera $x>1$, entonces $x^2>x$, y, por tanto, no sería x el máximo. En consecuencia, x debe ser igual a 1.

nectándola en dos puntos con los dos polos de una pila eléctrica. No cabe duda de que esta experiencia física conduce a un resultado definido. Pero ¿qué ocurre con el problema matemático correspondiente, que es de la máxima importancia en teoría de funciones y en otros campos? De acuerdo con la teoría de la electricidad, el fenómeno físico se describe como un «problema de valores de contorno de una ecuación en derivadas parciales». Es el problema matemático el que nos interesa; su resolubilidad parece posible, puesto que equivale a un fenómeno físico; pero esta argumentación no prueba nada matemáticamente. Riemann trató la cuestión matemática en dos etapas. Primero, demostró que el problema equivale a un problema de mínimo: cierta cantidad que expresa la energía del flujo eléctrico es mínima en el caso del flujo real en comparación con los otros flujos posibles, dentro de las condiciones prescritas. Enunció, después, como *principio de Dirichlet*, que tal problema de mínimo tiene solución. Riemann no dió el menor paso en busca de una demostración matemática de la segunda afirmación, y éste fué el punto objeto de los ataques de Weierstrass. No sólo no era evidente la existencia de un mínimo, sino que además resultó ser una cuestión sumamente delicada, para cuyo estudio no estaba preparada la matemática de aquella época, y que, finalmente, sólo se resolvió después de muchas décadas de investigación intensa.

2. **Ejemplos.**—Explicaremos la índole de la dificultad a que se hace mención, mediante dos ejemplos: 1) Sobre una recta L tomemos dos puntos A y B a una distancia d y busquemos la poligonal de

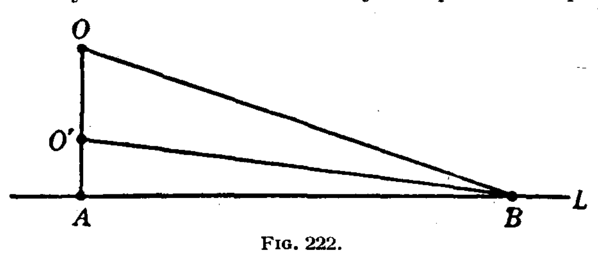


Fig. 222.

longitud mínima que partiendo de A en dirección perpendicular a L, termina en B. Puesto que el segmento rectilíneo AB es la distancia más corta de A a B entre todos los caminos posibles, podemos estar seguros que cualquier otro camino admisible tiene una longitud mayor que d, pues el único que tiene esa misma longitud d es el segmento rectilíneo AB, el cual no cumple la restricción impuesta en cuanto a la dirección en A, y, por consiguiente, no es admisible dentro de las

condiciones del problema. Por otra parte, si se considera el camino admisible AOB (Fig. 222) se encuentra que, al reemplazar O por otro punto O' suficientemente próximo al A, podemos obtener un camino admisible, cuya longitud diferirá tan poco como se quiera de d; en consecuencia, si existe un camino admisible de *longitud mínima*, ésta no puede ser mayor que d, por lo que debe ser exactamente igual a d. Pero el único camino que tiene tal longitud no es admisible; por tanto, no puede existir un camino de longitud mínima cumpliendo dicha condición, y el problema propuesto no tiene solución.

2) Sea C una circunferencia (Fig. 223) y S un punto a distancia l por encima de su centro; consideremos el conjunto de todas las superficies limitadas por C, que pasan por el punto S y que se encuentran por encima de C, de tal forma que dos puntos distintos no tengan la misma proyección vertical sobre el plano de C. ¿Cuál es, entre todas estas superficies, la de área mínima? Por muy natural que parezca este problema, carece de solución: no existe una superficie que cumpla los requisitos del problema y tenga área mínima. Si no se exige que la superficie pase por S, la solución será evidentemente el disco circular plano limitado por C, y cuya área llamaremos A. Cualquier otra superficie limitada por C, debe tener área mayor que A. Pero podemos encontrar una superficie que cumpla las condiciones del problema y cuya área exceda de A tan poco como queramos. Para esto consideremos una superficie cónica de altura l y de radio tan pequeño en la base que su área sea menor que cualquier valor asignado. Coloquemos este cono sobre el disco de tal modo que su vértice coincida con S y consideremos la superficie total formada por la lateral del cono y la parte del disco que no está cubierta por él. Es evidente que esta superficie, la cual sólo se desvía del disco en las proximidades de su centro, tiene un área que excede de A en menos del valor asignado, y como éste puede elegirse arbitrariamente pequeño, resulta que el mínimo, si existe, no puede ser sino el área A del disco. Pero, entre todas las superficies limitadas por C, sólo el disco tiene esta área, y como no pasa por S, no cumple una de las condiciones del problema. En consecuencia, éste carece de solución.

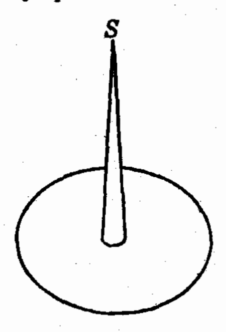


FIG. 223.

Podemos prescindir de otros ejemplos más complicados que dió Weierstrass; los dos considerados bastan para probar que la existencia de un mínimo no es una parte trivial de una demostración

matemática. Expongamos la cuestión de una manera más general y abstracta. Consideremos una clase definida de entes, p. ej., curvas o superficies, a cada una de las cuales se le asigna como función una longitud o un área. Si la clase se compone sólo de un número finito de objetos, entre los números correspondientes debe existir evidentemente un máximo y un mínimo. Pero si la clase se compone de infinitos objetos, no existe necesariamente ni lo uno ni lo otro, ni siquiera en el caso en que estos números estén comprendidos entre límites fijos. En general, dichos números formarán un conjunto infinito de puntos sobre la recta numérica. Supongamos, para mayor sencillez, que todos son positivos. Entonces, el conjunto tendrá un «extremo inferior»; es decir, un punto α a cuya izquierda no se encuentra ningún elemento del conjunto y que es o un elemento del mismo o tal que los elementos del conjunto se aproximan a él tanto como se desee. Si α pertenece al conjunto es su elemento mínimo; en otro caso, el conjunto carece de mínimo. Así, p. ej., el conjunto de los números 1, 1/2, 1/3, ..., no contiene elemento mínimo, pues el extremo inferior, 0, no pertenece al conjunto. Estos ejemplos aclaran de una manera abstracta las dificultades lógicas relacionadas con el problema de existencia. La solución matemática de un problema de mínimo no queda completa hasta haber demostrado, explícita o implícitamente, que el conjunto de valores asociado al problema contiene un elemento menor que todos los restantes.

3. Problemas elementales de extremos.—En los problemas elementales, sólo se requiere un análisis atento de los conceptos fundamentales en que se basan para decidir la cuestión de la existencia de una solución. En el capítulo VI se discutió el concepto general de conjunto compacto, y vimos allí que una función continua, definida para los elementos de un conjunto compacto, siempre alcanza un máximo y un mínimo en algún punto del conjunto. En cada uno de los problemas elementales que hemos estudiado, los valores que se discuten pueden considerarse como los de una función de una o varias variables definida en un dominio, que es un conjunto compacto o puede reducirse a él sin modificación esencial del problema. En tal caso, está asegurada la existencia de máximo o de mínimo; en el problema de Steiner, p. ej., la cantidad considerada era la suma de tres distancias, que depende de forma continua de la posición del punto móvil. Puesto que el dominio de este punto abarca todo el plano, nada se pierde si encerramos la figura en un círculo de gran radio y nos restringimos a que el punto varíe en su interior y sobre la circunferencia. Pues tan pronto como el punto móvil se encuentra

suficientemente alejado de los tres puntos dados, la suma de sus distancias a éstos será mayor que $AB + BC$, que es uno de los valores posibles de la función. En consecuencia, si existe un mínimo para un punto dentro del círculo, éste será también el mínimo para el problema sin aquella restricción. Pero es fácil probar que el dominio compuesto por una circunferencia y su interior es compacto; por tanto, existe un mínimo para el problema de Steiner.

La importancia de suponer que el dominio de variación de la variable independiente es compacto se pone de manifiesto mediante el siguiente ejemplo: dadas dos curvas cerradas, C_1 y C_2, existen siempre dos puntos, P_1 y P_2, situados, respectivamente, sobre C_1 y C_2 tales que su distancia es mínima; así como otros dos puntos, Q_1 y Q_2, cuya

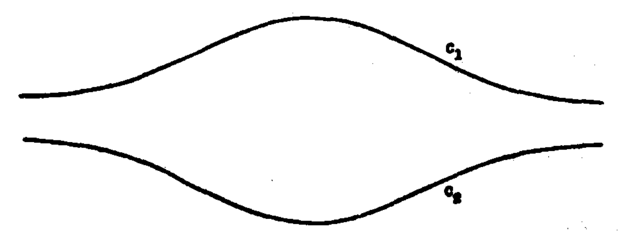


FIG. 224.—Curvas entre las que no hay distancia máxima ni mínima.

distancia es máxima, pues la distancia entre un punto A_1 de C_1 y otro A_2 de C_2 es una función continua en el conjunto compacto formado por las parejas de puntos A_1 y A_2 que se consideran. Sin embargo, si las dos curvas no están limitadas, sino que se extienden hasta el infinito, el problema puede carecer de solución. En el caso expuesto en la figura 224, no existe ni mínimo ni máximo de la distancia entre las dos curvas; el extremo inferior de aquélla es 0, y el superior, infinito, no alcanzándose ninguno de los dos. En algunos casos, existe mínimo pero no máximo. Para el caso de las dos ramas de la hipérbola (Fig. 17) sólo se alcanza la distancia mínima, en A y A', puesto que evidentemente no existen dos puntos separados por una distancia máxima.

Podemos explicar este diferente comportamiento, restringiendo artificialmente el dominio de las variables. Elijamos un número positivo arbitrario R y restrinjamos la x por la condición $|x| \leq R$. Entonces existe tanto un máximo como un mínimo para los dos últimos problemas. En el primero la limitación del contorno en esta forma asegura la existencia de una distancia máxima y otra mínima, que se alcanzan ambas en la frontera. Si R aumenta, los puntos en que se alcanzan los extremos se encuentran también sobre el contorno. En

consecuencia, al aumentar R, esos puntos se alejan infinitamente. En el segundo caso, la distancia mínima se alcanza en el interior, y por mucho que se aumente R, los dos puntos cuya distancia es mínima siguen siendo los mismos.

4. Dificultades en casos más complicados.—Mientras que la cuestión de existencia no es muy grave en los problemas elementales de una, dos o a lo sumo un número finito de variables independientes, cambia el aspecto por completo cuando se trata del principio de Dirichlet e incluso en casos más simples de tipo análogo. La razón de ello es que o bien el dominio de la variable independiente no es com-

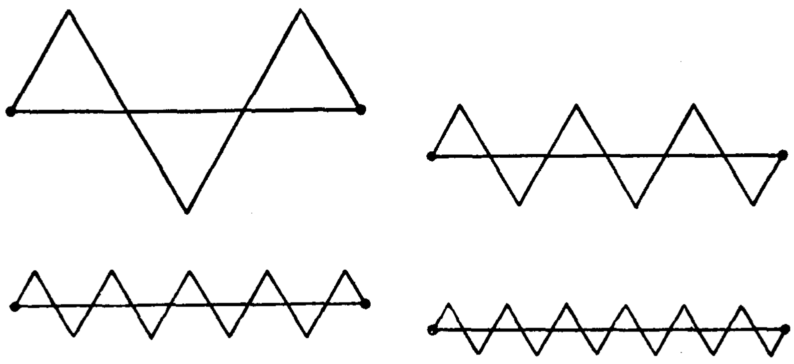


FIG. 225. — Aproximación a la longitud del segmento por polígonos de longitud doble.

pacto, o la propia función no es continua. En el ejemplo tratado en la página 378, teníamos una sucesión de caminos $AO'B$ tal que O' tendía al punto A. Cada camino de la sucesión satisface las condiciones impuestas, pero los caminos $AO'B$ tienden al segmento rectilíneo AB, el cual no pertenece al conjunto de las trayectorias permitidas. El conjunto de los caminos admisibles es, en este aspecto, análogo al intervalo $0 < x \leqslant 1$, para el cual no es válido el teorema de Weierstrass sobre valores extremos (pág. 324). En el segundo ejemplo nos encontramos en una situación análoga: si disminuye cada vez más el radio de la base de los conos, la sucesión de las correspondientes superficies admisibles tenderá al disco circular, más un segmento rectilíneo vertical que alcanza hasta S. Este ente geométrico, sin embargo, no figura entre las superficies admisibles, y ocurre nuevamente que el conjunto de dichas superficies no es compacto.

Como ejemplo de una dependencia no continua tenemos la longitud de una curva. Dicha longitud no es una función de un número finito de variables, puesto que una curva no puede caracterizarse por

un número finito de «coordenadas», y no es una función continua de la curva. Para comprender bien esto, unamos dos puntos A y B, situados a una distancia d, mediante una poligonal P_n, que junto con el segmento AB forma n triángulos equiláteros. Se deduce claramente de la figura 225 que la longitud total de P_n será exactamente $2d$, cualquiera que sea el valor de n. Consideremos ahora la sucesión de poligonales P_1, P_2, ... La altura de cada una de las ondas disminuye a medida que crece su número, y es evidente que la poligonal P_n tiende al segmento AB. La longitud de P_n es siempre $2d$, cualquiera que sea el subíndice n, mientras que la longitud de la curva límite, el segmento rectilíneo, sólo es d. De ahí que la longitud no dependa de la curva de manera continua.

Todos estos ejemplos confirman que las precauciones respecto de la existencia de solución son realmente necesarias en los problemas de mínimo de una estructura más compleja.

VIII. EL PROBLEMA DE LOS ISOPERÍMETROS

Uno de los hechos *evidentes* de la matemática, para el cual sólo los métodos modernos han dado una demostración rigurosa, es el referente a que, de todas las curvas cerradas de igual longitud, la circunferencia es la que encierra un área mayor. Steiner encontró diversas demostraciones ingeniosas de este teorema, de las cuales sólo consideraremos una.

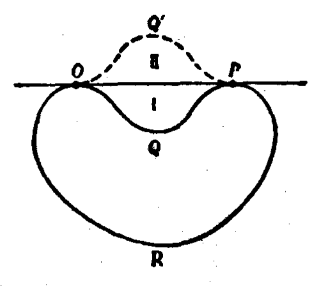


Fig. 226.

Comencemos por suponer que existe solución. Concedido esto, supongamos que la curva C es la requerida; o sea, tiene la longitud prescrita y encierra el área máxima. Es fácil ver que C debe ser convexa; es decir, todo segmento rectilíneo que una dos puntos cualesquiera de ella, debe estar situado enteramente dentro de C o sobre C. En efecto, si C no fuera convexa, como en la figura 226, sería posible trazar un segmento, tal como OP, entre dos puntos O y P de C, situado fuera de C; el arco $OQ'P$, simétrico de OQP respecto a la recta OP, forma junto con el arco ORP una curva de longitud L, que encierra una superficie mayor que C, puesto que incluye además las áreas adicionales I y II. Esto contradice la hipótesis según la cual C contiene el área máxima entre todas las curvas cerradas de longitud L. Por tanto, C debe ser convexa.

Elijamos ahora dos puntos, A, B, que dividan la curva solución C en dos arcos de igual longitud. Entonces, la recta AB debe dividir el área de C en dos partes iguales, pues en otro caso, hallando la simétrica respecto a AB de la mayor de las dos (Fig. 227) se obtendría otra curva de longitud L, que encerraría un área mayor que la de C. Se deduce de ahí que la mitad de la solución C debe resolver el siguiente problema: encontrar un arco de longitud $L/2$, cuyos extremos A y B estén sobre una recta dada y tal que se encierre un área máxima entre dicho arco y la recta. Probaremos ahora que la solución de este nuevo problema es una semicircunferencia, por lo que la curva C que resuelve el primer problema es una circunferencia.

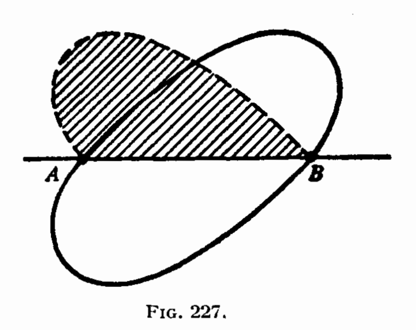


FIG. 227.

Sea el arco AOB la solución del nuevo problema. Basta demostrar que todo ángulo inscrito AOB (Fig. 228) es recto, con lo que quedará establecido que AOB es una semicircunferencia. Supongamos, por el contrario, que el ángulo AOB no fuera recto. Entonces, sustituímos la figura 228 por la 229, en la cual las áreas rayadas y la longitud del arco AOB no han variado, pero el área triangular ha aumentado al tomar el ángulo AOB igual o muy próximo a 90°. Así, la figura 229

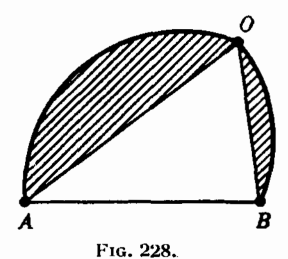


FIG. 228.

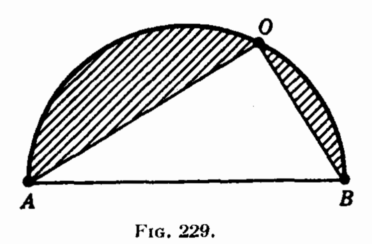


FIG. 229.

proporciona un área mayor que la original (véase pág. 341). Pero habíamos hecho la hipótesis de que la figura 228 resolvía el problema, de forma que la segunda figura no puede dar un área mayor. Esta contradicción prueba que para todo punto O el ángulo AOB debe ser recto, con lo que se completa la demostración.

Esta propiedad isoperimétrica del círculo puede expresarse mediante una desigualdad. Si L es la circunferencia del círculo, su área

será $L^2/4\pi$, por lo que debe existir la *desigualdad isoperimétrica* $A \leqslant L^2/4\pi$ entre el área A y la longitud L de cualquier curva cerrada, verificándose el signo de igualdad sólo para el círculo.

*Como se deduce de la discusión hecha en las páginas 376-82, la demostración de Steiner tiene sólo un valor condicional: «Si existe una curva de longitud L, de área máxima, ha de ser una circunferencia.» Para demostrar esta premisa hipotética se necesita una nueva argumentación. Probemos primero un teorema elemental, que se refiere a los polígonos cerrados P_n, con un número par de lados, $2n$, el cual dice que entre todos los polígonos de $2n$ lados y del mismo perímetro, el $2n$-ágono regular encierra la superficie máxima. La demostración coincide en lo fundamental con el razonamiento de Steiner, si bien contiene algunas modificaciones. Aquí no se plantea ninguna dificultad respecto al problema de existencia, puesto que el polígono de $2n$ lados, así como su perímetro y su área, dependen de manera continua de las $4n$ coordenadas de sus vértices, que sin pérdida de generalidad pueden quedar restringidas a un conjunto compacto de puntos en un espacio de $4n$ dimensiones. De acuerdo con ello, en este problema de polígonos podemos empezar suponiendo que cierto polígono P *es* la solución y, partiendo de esta base, analizar las propiedades de P. Exactamente como en la demostración de Steiner, se deduce que P debe ser convexo. Demostraremos ahora que los $2n$ lados de P tienen la misma longitud. Supongamos que dos lados adyacentes, AB y BC, tuvieran longitud diferente. Entonces podríamos separar el triángulo ABC de P y reemplazarlo por otro triángulo isósceles $AB'C$, en el cual $AB'+B'C=$ $= AB + BC$ y que tiene mayor área (véase pág. 344). Así obtendríamos un polígono P' con el mismo perímetro y área mayor, lo que contradice la hipótesis según la cual P era el polígono de área máxima de $2n$ lados. En consecuencia, todos los lados de P son de igual longitud, quedando por demostrar que es regular, para lo que basta probar que todos los vértices de P están sobre una circunferencia. El razonamiento sigue también paralelo al de Steiner. Previamente, demostraremos que toda diagonal que une dos vértices opuestos (p. ej., el primero con el $n+1$) divide la superficie en dos partes iguales, y a continuación probaremos que todos los vértices de cada una de esas partes están sobre una semicircunferencia. Los detalles, que siguen exactamente la marcha anterior, quedan como ejercicio a cargo del lector.

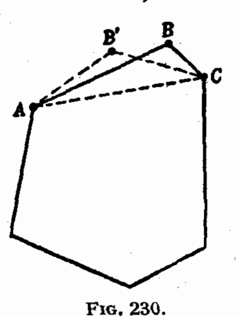


FIG. 230.

Puede probarse ahora la existencia, así como la solución del problema, mediante un proceso de límite, en el cual tiende a infinito el número de vértices y el polígono regular óptimo tiende a un círculo.

El razonamiento de Steiner no es completamente adecuado para probar la correspondiente propiedad isoperimétrica de la esfera en tres dimensiones. El mismo Steiner dió una argumentación más complicada y algo distinta, que resulta adecuada lo mismo para tres dimensiones que para dos; pero, dado que no puede adaptarse fácilmente de manera que proporcione la demostración de la existencia, preferimos omitirla. En efecto, la demostración de la propiedad isoperimétrica de la esfera es tarea mucho más difícil que la del círculo. Sólo mucho más tarde H. A. Schwarz dió por primera vez una demostración completa y rigurosa en una memoria bastante abstrusa. La propiedad isoperimétrica tridimensional puede expresarse mediante la desigualdad $36\pi V^2 \leqslant A^3$ entre el área A y el volumen V de cualquier cuerpo cerrado tridimensional, verificándose el signo de igualdad sólo para la esfera.

*IX. PROBLEMAS DE EXTREMOS CON CONDICIONES DE CONTORNO. RELACIÓN ENTRE EL PROBLEMA DE STEINER Y EL DE LOS ISOPERÍMETROS

Se obtienen resultados interesantes en los problemas de máximos y mínimos, cuando el dominio de la variable está restringido por condiciones de contorno. El teorema de Weierstrass, relativo a que una función continua en un dominio compacto alcanza un valor máximo y otro mínimo, no excluye la posibilidad de que alcance los valores extremos en el contorno del dominio. La función $u = x$ proporciona un ejemplo sencillo, casi trivial. Si x no está acotada y puede tomar cualquier valor entre $-\infty$ y $+\infty$, el dominio B de la variable independiente es toda la recta numérica, por lo que la función $u = x$ no alcanza en ningún punto un máximo o un mínimo. Pero si el dominio B está limitado y se reduce, p. ej., a $0 \leqslant x \leqslant 1$, existe un máximo, 1, que la función alcanza en su extremo derecho, y un mínimo, 0, que la función alcanza en el extremo izquierdo. Sin embargo, ninguno de estos valores extremos está representado por una cumbre o una depresión en la gráfica de la función; no son extremos relativos respecto a un entorno completo. Cambian en cuanto se extiende el intervalo, puesto que siempre permanecen en los puntos extremos. En cambio, una verdadera «cumbre» o «depresión» de una función tiene siempre un carácter de máximo o mínimo respecto a un entorno completo del punto donde la alcanza y no queda afectada por ligeros cambios en

los límites. Un extremo de este tipo persiste, aun cuando varíe libremente la variable independiente en el dominio *B*, por lo menos en un entorno suficientemente pequeño. La distinción entre esos extremos relativos y los que alcanza en el contorno queda aclarada por numerosos contextos, aparentemente inconexos. Para las funciones de una variable, naturalmente, la distinción es la que existe entre las funciones monótonas y las que no lo son, lo que no conduce a ninguna observación particularmente interesante; pero existen numerosos ejemplos muy significativos de valores extremos alcanzados en el contorno del dominio de variabilidad, en el caso de las funciones de varias variables.

Esto puede ocurrir, p. ej., en el problema del triángulo de Schwarz; en este caso, el dominio de variabilidad de las tres variables independientes consta de todas las ternas de puntos, una en cada lado del triángulo *ABC*. La solución del problema entraña dos posibilidades: o se llega al mínimo, cuando cada uno de los tres puntos, *P*, *Q*, *R*, que varían independientemente, se encuentra dentro de los respectivos lados del triángulo, en cuyo caso el mínimo está dado por el triángulo órtico, o el mínimo se obtiene en la posición de contorno, cuando dos de los puntos *P*, *Q*, *R* coinciden con el extremo común de sus respectivos intervalos, en cuyo caso el «triángulo» mínimo inscrito es la altura correspondiente a este vértice contada dos veces. Así, el carácter de la solución es completamente distinto en uno y otro caso.

En el problema de Steiner de las tres ciudades, el dominio de variabilidad del punto *P* es todo el plano, del cual pueden considerarse como formando parte del contorno los puntos *A*, *B* y *C*. También aquí aparecen dos casos posibles, que proporcionan tipos enteramente distintos de solución. O se alcanza el mínimo en el interior del triángulo *ABC*, caso de los tres ángulos iguales, o se obtiene en un punto *C* del contorno. Existe un par de casos análogos para el problema complementario.

Como ejemplo final podemos considerar el problema de los isoperímetros, modificado mediante condiciones restrictivas de contorno. Obtendremos así una sorprendente relación entre este problema y el de Steiner, y al mismo tiempo el ejemplo quizá más sencillo de un nuevo tipo de problema de extremos. En el problema original, la variable independiente, o sea, la curva cerrada de longitud dada, podía desviarse arbitrariamente de la forma circular, y cualquier curva así deformada es admisible, por lo que obtenemos un auténtico mínimo libre o relativo. Examinemos ahora el siguiente problema modificado: las curvas *C* en consideración deberán contener en su inte-

rior o pasar por tres puntos dados, P, Q, R, estando dada el área A y debiendo hacerse mínima la longitud L. Esto representa una verdadera condición de contorno.

Es evidente que si el área dada A es suficientemente grande, los tres puntos P, Q, R no afectarán en nada al problema. Siempre que el círculo circunscrito al triángulo PQR tenga área menor o igual que A, la solución será simplemente un círculo de área A que incluya a los tres puntos. Pero, ¿qué ocurre si A es menor? Nos contentaremos con enunciar la solución omitiendo la demostración detallada,

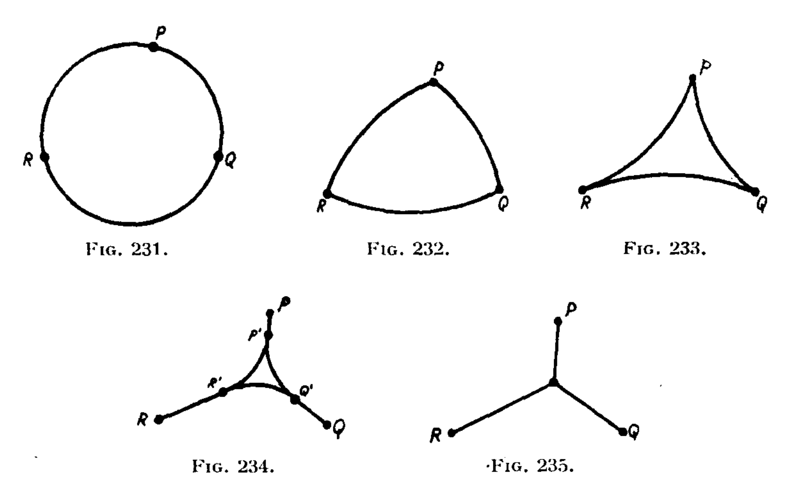


FIG. 231. FIG. 232. FIG. 233.

FIG. 234. FIG. 235.

FIGS. 231-35.—Figuras isoperimétricas que tienden a la solución del problema de Steiner.

aunque no está fuera de nuestras posibilidades. Caractericemos las soluciones mediante una sucesión de valores de A que tienden a cero. Apenas el valor de A es menor que el área del círculo circunscrito, el círculo isoperimétrico primitivo se descompone en tres arcos, todos del mismo radio, que forman un triángulo curvilíneo convexo de vértices P, Q, R (Fig. 232). Este triángulo es la solución y sus dimensiones pueden determinarse por el valor dado de A. Si A continúa disminuyendo, aumentará el radio de esos arcos, tendiendo a ser segmentos rectilíneos, hasta que A es exactamente igual al área del triángulo PQR, en cuyo caso la solución es este mismo triángulo. Si A sigue disminuyendo, la solución constará otra vez de tres arcos circulares, todos del mismo radio y que forman un triángulo de vértices P, Q, R (Fig. 233). Si A continúa disminuyendo, llegará un momento en el que, para un cierto valor de A, dos de los arcos cóncavos

serán mutuamente tangentes en un vértice R. Si A disminuye todavía más, no será ya posible construir un triángulo curvilíneo del tipo anterior. Se presenta un nuevo fenómeno: todavía la solución es un triángulo curvilíneo cóncavo, pero uno de los vértices R' se ha separado del correspondiente vértice R, y la solución es ahora el triángulo PQR' más el segmento RR' contado dos veces (puesto que pasa de R a R' y de R' a R). Este segmento rectilíneo es tangente a los dos arcos, que son a su vez tangentes entre sí en R'. Si A sigue decreciendo, este fenómeno de separación se presentará también en los otros vértices. Eventualmente, se obtiene como solución un triángulo curvilíneo compuesto por tres arcos circulares de igual radio, tangentes entre sí y que forman un triángulo curvilíneo equilátero $P'Q'R'$, además de tres segmentos rectilíneos, contados dos veces cada uno, $P'P$, $Q'Q$, $R'R$ (Fig. 234). Si, finalmente, A se hace cero, el triángulo curvilíneo se reduce a un punto, volviendo a la solución del problema de Steiner. Se ve así que esta última circunstancia es el caso límite del problema modificado de los isoperímetros.

Si P, Q, R forman un triángulo obtuso con un ángulo mayor de 120°, el proceso anterior conduce a la solución del problema de Steiner, pues los arcos circulares tienden hacia el vértice del ángulo obtuso. Mediante procesos de límite de naturaleza análoga pueden obtenerse las soluciones del problema de Steiner generalizado (Figs. 216-218).

X. EL CÁLCULO DE VARIACIONES

1. Introducción.—El problema de los isoperímetros es sólo un ejemplo, probablemente el más antiguo de todos, de un tipo muy amplio de problemas, sobre los que llamó la atención Johann Bernoulli en 1696. En *Acta Eruditorum*, el gran periódico científico de la época, propuso el siguiente problema de la braquistocrona: imagínese una partícula obligada a deslizarse sin rozamiento a lo largo de cierta curva, que une un punto, A, con otro situado más abajo, B. Si se permite que la partícula descienda exclusivamente bajo la acción de la gravedad, se pregunta qué curva deberá elegirse para que el tiempo empleado en el descenso sea mínimo. Es fácil ver que la partícula empleará diferentes intervalos de tiempo al variar las trayectorias. De ningún modo proporciona la línea recta la trayectoria de tiempo mínimo, así como tampoco constituye la solución un arco de circunferencia o cualquier otra curva elemental. Bernoulli se enorgulleció de haber encontrado una solución maravillosa, que no quiso publicar inmediatamente para incitar a los grandes matematicos de la época a

probar su habilidad en este nuevo tipo de cuestiones matemáticas. En particular, desafió a su hermano mayor Jacob, con quien estaba entonces profundamente enemistado y a quien públicamente había calificado de incompetente para resolver el problema. Los matemáticos reconocieron inmediatamente que el problema de la braquistocrona era de un carácter enteramente distinto. Mientras que, hasta entonces, en los problemas estudiados mediante el cálculo diferencial, la cantidad cuyo mínimo se buscaba dependía sólo de una o más variables numéricas, en este problema la cantidad que se consideraba, es decir, el tiempo necesario para el descenso, dependía de *toda la curva*, lo que constituye una diferencia esencial, que ponía el problema fuera del alcance del cálculo diferencial o de cualquier otro método conocido en aquella época.

La novedad del problema (al parecer no se comprendió claramente entonces que el de los isoperímetros es de la misma naturaleza) fas-

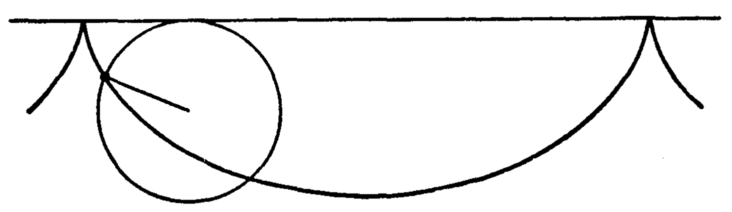


FIG. 236.—La cicloide.

cinó a los matemáticos contemporáneos, tanto más al conocerse que la solución la proporcionaba la cicloide, curva que había sido descubierta poco tiempo antes. (Recordemos la definición de la cicloide: es el lugar de un punto de una circunferencia que rueda sin deslizar sobre un recta, según se ve en la figura 236.) Esta curva había sido relacionada con algunos interesantes problemas de mecánica, especialmente con la construcción de un péndulo ideal. Huygens había descubierto que un punto ideal provisto de masa, que oscile sin rozamiento bajo la acción de la gravedad en una cicloide vertical, tiene un período de oscilación independiente de la amplitud del movimiento. En una trayectoria circular, como la de un péndulo ordinario, esa independencia es sólo aproximada, lo que se tenía por un inconveniente en la utilización del péndulo en los relojes de precisión. Se había honrado a la cicloide llamándola tautocrona; pero entonces se le asignó el nuevo título de braquistocrona.

2. El cálculo de variaciones. El principio de Fermat en óptica. Entre los diferentes métodos de resolver el problema de la braquis-

tocrona, que encontraron Bernoulli y otros investigadores, expondremos el más original. Los primeros métodos eran de un carácter más o menos especial, adaptados al problema particular en cuestión. Pero no pasó mucho tiempo sin que Euler y Lagrange (1736-1813) desarrollaran métodos más generales para resolver problemas de extremos, en los cuales el elemento independiente no era ya una sola variable numérica o un número finito de tales variables, sino toda una curva o función e incluso un sistema de funciones. Se llamó *cálculo de variaciones* al nuevo método que permitía resolver dichos problemas.

No es posible describir aquí los aspectos técnicos de esta rama de la matemática o profundizar en la discusión de problemas concretos. El cálculo de variaciones tiene numerosas aplicaciones en física y hace ya mucho tiempo que se observó que los fenómenos naturales se ajustan a menudo a algún principio de máximos y mínimos. Como ya hemos visto, Herón de Alejandría comprendió que podía describirse la reflexión de un rayo luminoso en un espejo plano mediante un principio de mínimo. En el siglo XVII, Fermat dió un paso más en esta dirección, al observar que la ley de la refracción de la luz puede expresarse también mediante un principio de mínimo.

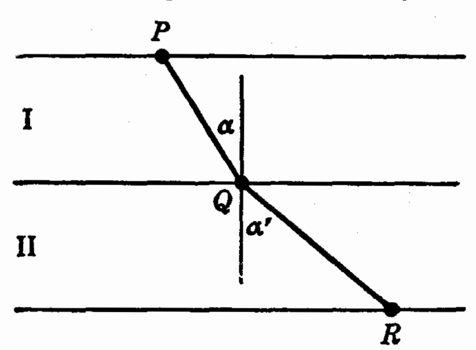


FIG. 237.—Refracción de un rayo luminoso.

Se sabe que la trayectoria de un rayo luminoso, que pasa de un medio homogéneo a otro de la misma naturaleza, se desvía en la superficie de separación de ambos. Así, en la figura 237, un rayo luminoso que parte de P en el medio superior donde la velocidad es v, hacia R, que se encuentra en el medio inferior donde la velocidad es w, seguirá la trayectoria PQR. La ley empírica encontrada por Snell (1591-1626) afirma que la trayectoria consta de dos segmentos rectilíneos, PQ y QR, que forman con la normal dos ángulos, α y α', determinados por la condición: sen α/sen $\alpha' = v/w$. Por medio del cálculo, Fermat demostró que esta trayectoria es tal que el tiempo necesario para que el rayo de luz pase de P a R es mínimo; es decir, menor de lo que sería si siguiera cualquier otra trayectoria. De esta forma, la ley de reflexión de Herón fué complementada 1600 años más tarde mediante otra, de refracción, análoga a la primera e igualmente importante.

Fermat generalizó esta ley de tal modo que incluyera superficies curvas de separación de los diferentes medios, como ocurre en las lentes esféricas. También en este caso se cumple la ley, pues la luz sigue una trayectoria tal que requiere un tiempo mínimo, comparado con el que sería necesario siguiendo otra trayectoria distinta entre los mismos dos puntos. Finalmente, Fermat consideró cualquier sistema óptico en el cual la velocidad de la luz varíe de manera determinada de un punto a otro, como ocurre en la atmósfera. Dividió el medio continuo, pero no homogéneo, en capas paralelas, en cada una de las cuales la velocidad de la luz es aproximadamente constante, e imaginó que este medio quedaba reemplazado por otro, en el cual la velocidad en cada capa es realmente constante. Entonces pudo aplicar nuevamente su principio, pasando de una capa a otra. Haciendo que el espesor de las capas tendiera a cero, obtuvo *el principio general de Fermat de la óptica geométrica*: en un medio no homogéneo, un rayo luminoso que pasa de un punto a otro sigue aquella trayectoria para la cual es mínimo el tiempo respecto al que requiere cualquier otra que una los mismos dos puntos. Este principio ha tenido una importante trascendencia, no sólo desde el punto de vista teórico, sino en la práctica de la óptica geométrica. Aplicando a este principio las técnicas del cálculo de variaciones se obtienen las bases para calcular los sistemas de lentes.

En otras ramas de la física los principios de mínimo han adquirido también una importancia enorme. Se observó que se obtiene el equilibrio estable de un sistema mecánico si se dispone de tal forma que su «energía potencial» sea mínima. Como ejemplo, consideraremos una cadena homogénea y flexible, suspendida por sus extremos y sobre la que actúa libremente la fuerza de la gravedad. La cadena adoptará aquella forma según la cual su energía potencial es mínima. En este caso, la energía potencial está determinada por la altura del centro de gravedad respecto a cierto eje fijo. La curva que adopta la cadena en suspensión se llama catenaria, y tiene cierto parecido con la parábola.

No sólo las leyes del equilibrio, sino también las del movimiento, están dominadas por principios de máximo o mínimo. Fué Euler el primero que tuvo ideas claras acerca de estos principios, aunque algunos temperamentos, inclinados hacia la especulación mística y filosófica, como Maupertius (1698-1759), no pudieron separar las proposiciones matemáticas de ciertas ideas peregrinas acerca de la «intención de Dios de regular los fenómenos físicos mediante un principio general de la más alta perfección». Los principios físicos variacionales de Euler, redescubiertos y ampliados por el matemático irlandés W. R. Hamilton (1805-1865), han demostrado ser poderosísimas herramientas

en mecánica, óptica y electrodinámica, con numerosas aplicaciones a la ingeniería. El desarrollo reciente de la física—relatividad y teoría de los cuantos—está lleno de ejemplos que revelan el poder del cálculo de variaciones.

3. El método de Bernoulli y el problema de la braquistocrona. El primer método desarrollado por Jacob Bernoulli para tratar el problema de la braquistocrona puede comprenderse sin poseer grandes conocimientos técnicos. Se sabe por mecánica que si un grave cae desde A sin velocidad inicial, siguiendo una curva C, tendrá en cualquier punto P una velocidad proporcional a $\sqrt{h}$, siendo h la distancia vertical entre A y P; es decir, $v = c\sqrt{h}$, donde c es una constante. Reemplacemos ahora el problema dado por otro ligeramente distinto. Dividamos el espacio en numerosas y delgadas capas horizontales, cada una de espesor d, y supongamos por un momento que la velocidad de la partícula en movimiento no varía con continuidad, sino por pequeños saltos, al pasar de una capa a otra; es decir, que en la primera capa, adyacente a A, la velocidad es $c\sqrt{d}$; en la segunda, $c\sqrt{2d}$, y en la n-ésima, $c\sqrt{nd} = c\sqrt{h}$, siendo h la distancia vertical de A a P (Fig. 238). Si nos limitamos a este problema, sólo existe en realidad un número finito de variables. Al ser rectilínea la trayectoria en cada capa, no se plantea problema de existencia; la solución debe ser una poligonal, y la única dificultad consiste en determinar sus vértices. De acuerdo con el principio de mínimo de la ley de la refracción simple, en cada par de capas sucesivas el movimiento de P a R pasando por Q debe ser tal que, suponiendo P y R fijos, Q proporciona la curva de tiempo mínimo. Por tanto, puede aplicarse la siguiente «ley de refracción»:

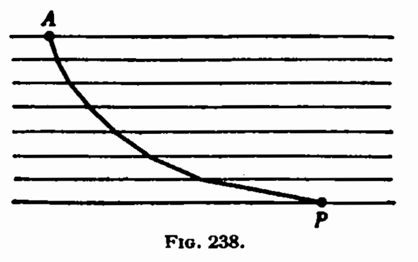


FIG. 238.

$$\frac{\operatorname{sen} \alpha}{\sqrt{nd}} = \frac{\operatorname{sen} \alpha'}{\sqrt{(n+1)d}}$$

Por aplicación reiterada de este razonamiento se obtiene la siguiente sucesión de igualdades:

$$\frac{\operatorname{sen} \alpha_1}{\sqrt{d}} = \frac{\operatorname{sen} \alpha_2}{\sqrt{2d}} = \cdots, \tag{1}$$

donde α_n es el ángulo entre la poligonal en la n-ésima capa y la vertical.

Bernoulli supone ahora que el espesor d se hace cada vez más pequeño y tiende a cero, de manera que el polígono que se acaba de obtener como solución del problema aproximado tiende a la solución buscada del problema original. Este paso al límite no afecta a las igualdades [1], por lo que Bernoulli concluye que la solución debe ser una curva C, que goce de la siguiente propiedad: si α es el ángulo formado por la tangente y la vertical en cualquier punto P de C, y h es la distancia vertical de P a la horizontal que pasa por A, sen $\alpha/\sqrt{h}$ es constante para todos los puntos P de C. Es fácil demostrar que esta propiedad caracteriza a la cicloide.

La «demostración» de Bernoulli es un ejemplo típico de una clase de razonamiento matemático, ingenioso y no exento de valor, pero que, al propio tiempo, carece de rigor. Existen varias suposiciones tácitas en la argumentación cuya justificación sería más complicada y engorrosa que la propia demostración; p. ej., se ha supuesto que existía una solución C y que la solución del problema aproximado tendía a la verdadera. Ciertamente, merece una atenta discusión el valor intrínseco de las consideraciones heurísticas de este tipo, pero esto nos llevaría demasiado lejos.

4. Geodésicas en una esfera. Geodésicas y maxi-mínimos.—En la introducción de este capítulo mencionamos el problema de determinar el arco mínimo que une dos puntos dados de una superficie. Sobre una esfera, se demuestra en geometría elemental que esas «geodésicas» son arcos de círculos máximos. Sean P y Q dos puntos (no diametralmente opuestos) de una esfera, y c el menor de los arcos del círculo máximo que los une. Se plantea entonces la cuestión de saber qué representa el arco mayor c' de este círculo máximo. Ciertamente, no proporciona ni la longitud máxima ni la mínima entre todas las curvas que unen P con Q, puesto que pueden trazarse entre ambos puntos curvas de longitud arbitrariamente grande. La respuesta es que c' resuelve un problema de maxi-mínimo. Consideremos un punto S sobre un círculo máximo que separe a P y Q; se trata de determinar la curva esférica

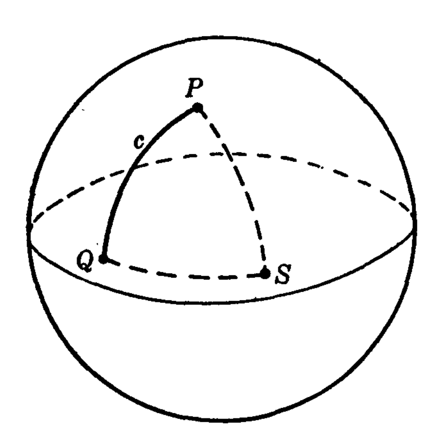


FIG. 239.—Geodésicas en una esfera.

de longitud mínima que une P con Q pasando por S. Naturalmente, el mínimo viene dado por una curva formada por dos pequeños arcos de círculo máximo, PS y QS. Busquemos ahora una posición del punto S de forma que esta distancia mínima PSQ sea la mayor posible. La solución es la siguiente: S debe ser tal que PSQ sea el arco mayor c' del círculo máximo PQ. Podemos modificar el problema buscando primero la trayectoria de longitud mínima que une P con Q pasando por n puntos prefijados S_1, S_2, ..., S_n de la esfera y tratar después de determinar los puntos S_1, S_2, ..., S_n de tal manera que esta distancia mínima sea la mayor posible. La solución viene dada por una trayectoria sobre el círculo máximo que une P con Q, pero que se arrolla tantas veces alrededor de la esfera que pasa exactamente n veces por los puntos diametralmente opuestos a P y Q.

Este ejemplo de problema de maxi-mínimo es típico de una clase muy amplia de cuestiones del cálculo de variaciones que han sido estudiadas con gran éxito mediante métodos desarrollados por Morse y otros autores.

XI. SOLUCIÓN EXPERIMENTAL DE PROBLEMAS DE MÍNIMO. EXPERIMENTOS CON PELÍCULAS

1. Introducción.—Generalmente es muy difícil, y a veces imposible, resolver explícitamente los problemas variacionales mediante fórmulas o construcciones geométricas en función de elementos simples ya conocidos. En lugar de ello, debemos conformarnos muchas veces con demostrar la existencia de una solución bajo determinadas condiciones e investigar después sus propiedades. En múltiples casos, cuando dicha demostración de existencia resulta ser más o menos dificultosa, es interesante estudiar las condiciones matemáticas del problema mediante ciertos artificios físicos o, mejor dicho, considerar el problema matemático como una interpretación de un fenómeno físico. La existencia de este último representará entonces la solución del problema matemático. Naturalmente, esto sólo constituye una justificación plausible sin llegar a ser una demostración matemática, pues queda todavía una cuestión por dilucidar, a saber: si la interpretación matemática del hecho físico es adecuada en sentido estricto, o si proporciona sólo una imagen poco apropiada de la realidad física. Algunas veces tales experimentos, aunque efectuados sólo en la imaginación, son convincentes incluso para los matemáticos. En el siglo XIX, Riemann descubrió muchos de los teoremas fundamentales de la teoría de funciones razonando acerca de sencillos experi-

mentos relacionados con el flujo de la electricidad en láminas metálicas delgadas.

En esta sección vamos a discutir, basándonos en demostraciones experimentales, uno de los problemas más profundos del cálculo de variaciones; el llamado problema de Plateau, debido a que este físico belga (1801-1883) llevó a cabo interesantes experimentos sobre el particular. El problema en sí mismo es mucho más antiguo, remontándose su origen hasta las fases iniciales del cálculo de variaciones. En su forma más simple consiste en encontrar la superficie de área mínima limitada por un contorno cerrado dado en el espacio. Discutiremos también ciertos experimentos relacionados con cuestiones conexas, todo lo cual servirá para iluminar algunos de nuestros resultados anteriores, así como ciertos problemas matemáticos de un nuevo tipo.

2. Experimentos con soluciones jabonosas.—Matemáticamente, el problema de Plateau está relacionado con la solución de una *ecuación en derivadas parciales* o con un sistema de tales ecuaciones. Euler demostró que todas las superficies mínimas (no contenidas en un plano) deben tener forma de silla de montar, y que su curvatura media[1] en cada punto ha de ser cero. Durante el último siglo se probó la existencia de solución para numerosos casos especiales, pero dicha existencia sólo ha sido demostrada recientemente en el caso general por J. Douglas y T. Radò.

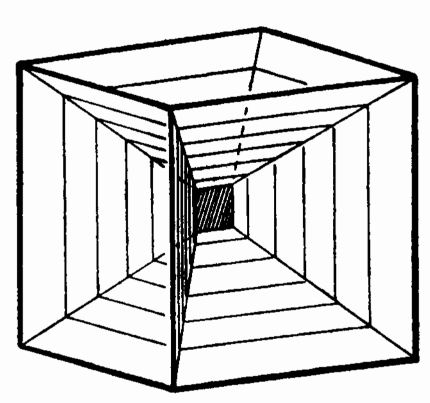


FIG. 240.—Estructura cúbica con una película jabonosa formada por 13 superficies casi planas.

Los experimentos de Plateau proporcionan inmediatamente soluciones para contornos muy generales. Si se sumerge cualquier contorno cerrado construído con alambre en un líquido de baja tensión superficial, al extraerlo, el contorno estará cubierto por una película que adopta la forma de una superficie de área mínima. (Se supone que puede despreciarse la gravitación y que lo mismo ocurre

[1] La curvatura media de una superficie en un punto P se define de la siguiente manera: consideremos la normal a la superficie en P y el haz de planos que pasan por ella. Estos planos cortan a la superficie según curvas que, en general, tienen curvatura diferente en P. Si consideramos las curvas de curvatura máxima y mínima, se llama curvatura media de la superficie P a la media aritmética de aquéllas (en general, los dos planos que dan las curvas de curvatura máxima y mínima son perpendiculares).

con otras fuerzas que se oponen a la tendencia de la película a adoptar una posición de equilibrio estable mediante la formación de una superficie de área mínima, que corresponde al menor valor posible de la energía potencial debida a la tensión superficial.) Una buena receta para preparar un líquido adecuado es la siguiente: disuélvanse 10 g de oleato sódico puro en 500 g de agua destilada y mézclense 15 unidades (en volumen) de esa solución con 11 de glicerina. Las películas obtenidas con esta solución sobre estructuras de alambre de cobre son relativamente estables. Las estructuras no deberán sobrepasar un diámetro total de 10 a 12 cm.

Con este método es muy fácil «resolver» el problema de Plateau, dando a la estructura de alambre la forma deseada. Se obtienen muy bellos modelos con estructuras de alambre de forma poligonal, formadas por una sucesión de aristas de un poliedro regular. En particular, es interesante sumergir una estructura cúbica completa en el líquido. Resulta primero un sistema de diferentes superficies que se cortan mutuamente, formando ángulos de 120°. (Si se retira cuidadosamente el cubo, existirán 13 superficies aproximadamente planas.) Después podemos perforar y destruir varias de estas superficies hasta que quede una sola limitada por un polígono cerrado. De esta manera pueden formarse bellísimas superficies. El mismo experimento puede hacerse con un tetraedro.

3. Nuevos experimentos sobre el problema de Plateau.—El propósito que se persigue con estos experimentos es mucho más amplio que el de las demostraciones originales de Plateau. En los últimos años se ha estudiado el problema de las superficies mínimas, no sólo con uno sino con un número cualquiera de contornos, siendo además la estructura topológica de la superficie mucho más complicada; p. ej., puede ser una superficie de una sola cara o de un género diferente de cero. Estos problemas más generales proporcionan una asombrosa variedad de fenómenos geométricos que pueden ponerse de manifiesto mediante experimentos con películas de solución jabonosa. A este respecto, es sumamente interesante efectuar los experimentos con estructuras de alambre flexible y estudiar el efecto de las deformaciones del contorno dado sobre la película.

Describiremos varios ejemplos:

1) Si el contorno es una circunferencia obtenemos un disco circular plano. Si deformamos continuamente el contorno, parece que la superficie mínima debería mantener siempre el carácter topológico de un disco, pero no ocurre así. Si se deforma el contorno hasta que adopte la forma indicada en la figura 241, obtenemos una superficie

mínima que ya no es simplemente conexa como el disco, sino que se ha convertido en una cinta de Moebius de una sola cara. Inversamente, podemos iniciar nuestro experimento con esta estructura y una pelí-

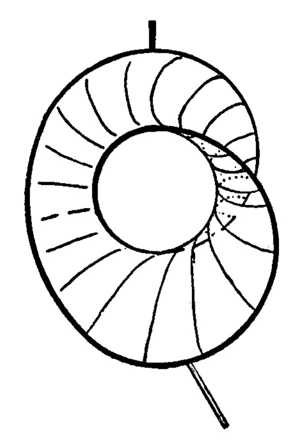


Fig. 241.—Cinta de Moebius de una sola cara.

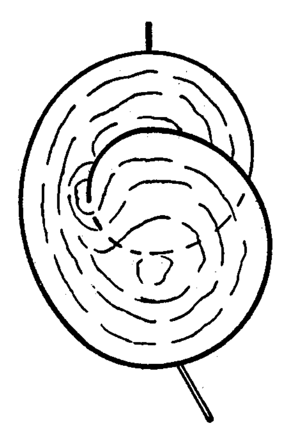


Fig. 242.—Superficie de dos caras.

cula de solución con la forma de una superficie de Moebius, y deformar la estructura mediante agarraderas soldadas a la misma (figura 241). En este experimento llega un momento en que cambia de repente el carácter topológico de la película, y la superficie adquiere nuevamente el tipo de un disco simplemente conexo (figura 242). Invirtiendo la deformación, se obtiene de nuevo una superficie de Moebius. En este experimento alternativo de deformación, la transformación de la superficie simplemente conexa en la cinta de Moebius ocurre en una etapa posterior. Esto prueba que debe existir toda una gama de diferentes formas del contorno, para las cuales la superficie de Moebius y la simplemente conexa son estables; es decir, proporcionan mínimos relativos. Pero cuando la superficie de Moebius tiene un área mucho menor que la otra, esta última resulta demasiado inestable para que pueda formarse.

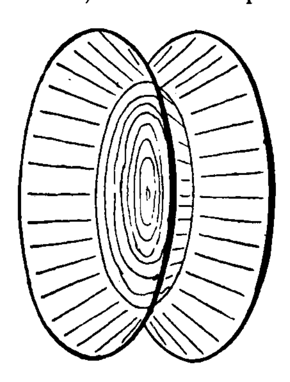


Fig. 243.—Sistema de tres superficies.

2) Podemos producir una superficie mínima de revolución entre dos circunferencias. Después de retirar la estructura de alambre de la solución, encontramos, no una superficie simple, sino tres superficies

que se cortan bajo ángulos de 120°, una de las cuales es un simple disco circular paralelo a los círculos del contorno (Fig. 243). Al destruir esta superficie intermedia, aparece la clásica catenoide [superficie que se obtiene por revolución de la catenaria (pág. 392) alrededor de una recta perpendicular a su eje de simetría]. Si se separan las dos circunferencias del contorno, llega un momento en que es inestable la superficie mínima doblemente conexa (catenoide), y en este momento se convierte en dos discos distintos. Naturalmente, el fenómeno no es reversible.

3) Otro ejemplo interesante lo proporcionan las estructuras de las figuras 244 a 246, sobre las cuales pueden extenderse tres superficies mínimas diferentes. Cada una está limitada por la misma curva sim-

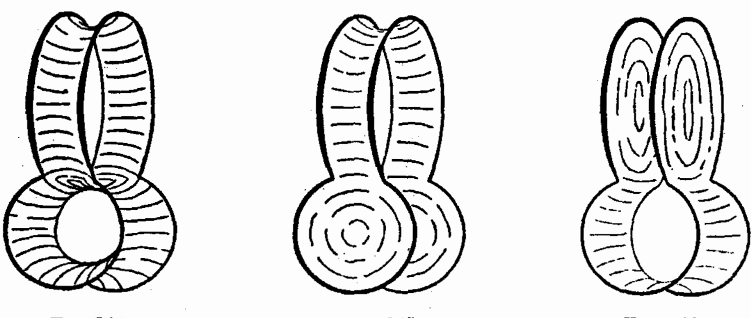


FIG. 244. FIG. 245. FIG. 246.
Estructura que produce tres superficies distintas, de géneros 0 y 1.

ple cerrada; una (Fig. 244) tiene género 1, mientras que las otras dos son simplemente conexas y, en cierto modo, simétricas. Estas dos últimas tienen igual área si el contorno es completamente simétrico. Pero si no es así, una de ellas proporciona el mínimo absoluto del área y la otra sólo da un mínimo relativo, siempre que el mínimo se busque entre las superficies simplemente conexas. La posibilidad de la solución de género 1 depende de que, al admitir superficies de ese tipo, es posible obtener áreas menores que en el caso en que se exija que la superficie sea simplemente conexa. Deformando radicalmente la estructura de alambre, debe llegar un momento en que esto ya no es posible; entonces, se hace cada vez mayor la inestabilidad de la superficie de género 1, transformándose repentina y discontinuamente en la superficie simplemente conexa y estable, representada en las figuras 245 y 246. Si iniciamos el experimento con una de esas superficies simplemente conexa (Fig. 246), podemos deformarla de tal manera que la otra superficie representada en la figura 245, que es también

simplemente conexa, sea mucho más estable. La consecuencia es que, en un cierto momento, se producirá una transición discontinua de la una a la otra. Invirtiendo lentamente la deformación, volvemos a la

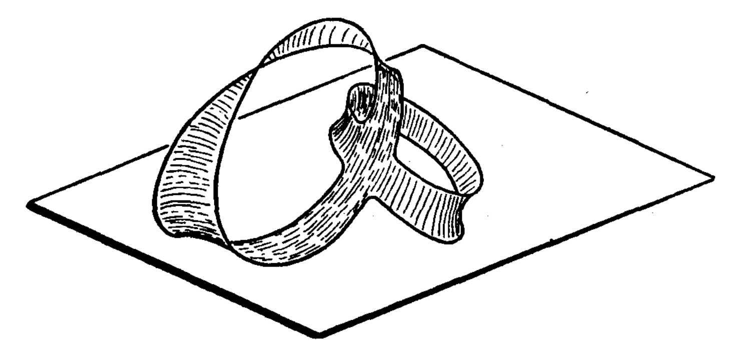


FIG. 247.—Superficie mínima de una sola cara, de estructura topológica superior y contorno único.

posición inicial de la estructura, aunque ahora aparece la otra superficie. Podemos repetir el fenómeno en dirección opuesta, obteniendo alternativamente, mediante transformaciones discontinuas, ambos tipos. Si se procede con cuidado, es posible transformar discontinuamente una de las dos soluciones simplemente conexa en la de género 1. Con este fin, debemos acercar todo lo posible las partes en forma de disco, con lo que la superficie de género 1 adquiere una notable estabilidad. Durante este experimento, ocurre algunas veces que aparecen primero películas intermedias de solución, que es necesario destruir antes que aparezca la superficie de género 1.

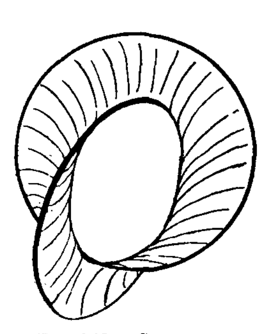


FIG. 248.—Curvas que se cruzan.

Este ejemplo muestra no sólo la posibilidad de la existencia de diferentes soluciones del mismo carácter topológico, sino también de otros tipos distintos en la misma estructura de alambre. Además, se deduce de nuevo la posibilidad de transformaciones discontinuas de una solución en otra, al variar con continuidad las condiciones del problema. Es fácil construir modelos más complicados de la misma clase y estudiar experimentalmente su comportamiento.

Un fenómeno interesante es la aparición de superficies mínimas limitadas por dos o más curvas cerradas que se cruzan. Para dos circunferencias se obtiene la superficie representada en la figura 248. Si,

en este ejemplo, se colocan los círculos de forma que sean perpendiculares entre sí y que la recta de intersección de sus planos sea un diámetro común, aparecerán dos formas simétricamente opuestas de esta superficie, con igual área. Si se varía ligeramente la posición mutua de ambos círculos, la superficie variará continuamente, si bien para cada posición existe sólo una superficie mínima absoluta, siendo la otra un mínimo relativo. Si se altera la posición de ambos círculos de tal modo que se forme el mínimo relativo, llegará un momento en que se transforme discontinuamente en el mínimo absoluto. En este caso ambas superficies mínimas posibles tienen el mismo carácter topológico, como las de las figuras 245 y 246, cada una de las cuales puede transformarse discontinuamente en la otra mediante una leve deformación de la estructura.

4. Soluciones experimentales de otros problemas matemáticos.—Debido a la acción de la tensión superficial, una película líquida sólo está en equilibrio estable cuando su área es mínima. Es ésta una fuente inagotable de experimentos, plenos de significado matemático. Si se permite que alguna parte del contorno de la película se mueva libre-

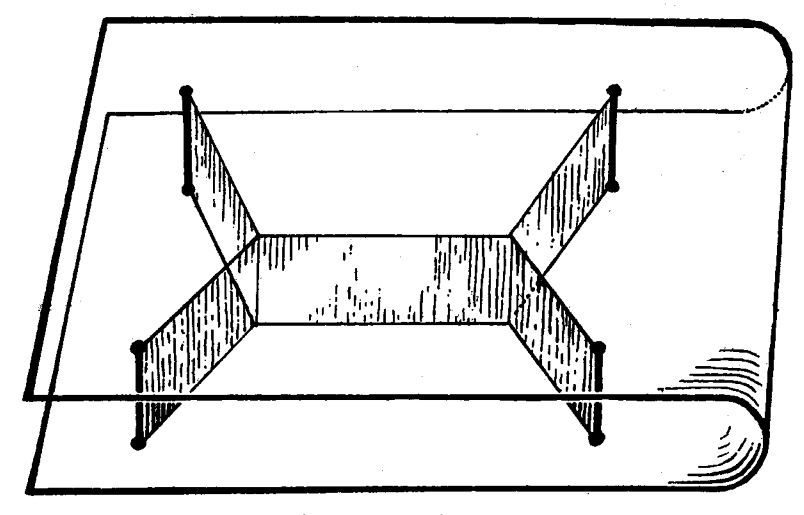


FIG. 249.—Demostración de la conexión mínima entre cuatro puntos.

mente sobre superficies dadas, la película será perpendicular en dicho contorno a la superficie prescrita.

Podemos utilizar este fenómeno para obtener cuiiosas demostraciones del problema de Steiner y sus generalizaciones (véanse páginas 364-71). Dos placas paralelas de vidrio o de algún material plás-

tico transparente están unidas mediante tres o más barras perpendiculares. Si sumergimos este artificio en una solución jabonosa, al extraerlo nuevamente veremos que la película forma un sistema de planos verticales entre las placas que unen las barras citadas. La proyección que aparece sobre las placas de vidrio constituye la solución del problema considerado en la página 369.

Si las placas no son paralelas, o las barras no son perpendiculares a las mismas, o bien se hace uso de placas curvadas, las curvas for-

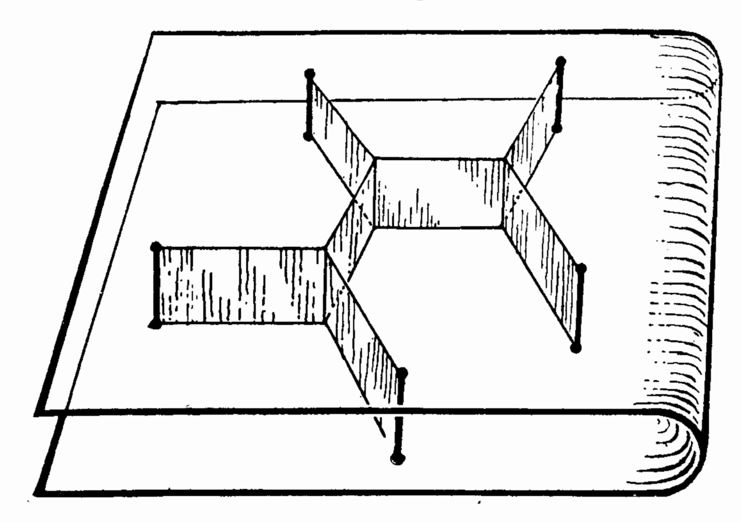


FIG. 250.—Conexión mínima entre cinco puntos.

madas por la película sobre las placas no serán ya rectilíneas, y constituyen ilustraciones de nuevos problemas variacionales.

El aspecto de la intersección de tres hojas de una superficie mínima que se cortan bajo ángulos de 120° puede considerarse como una generalización a más dimensiones de ciertos fenómenos relacionados con el problema de Steiner. Esto resulta evidente si, p. ej., unimos dos puntos del espacio, A y B, mediante tres curvas, y estudiamos el sistema estable correspondiente de películas de solución jabonosa. Como caso más sencillo consideremos que una de las curvas es el segmento rectilíneo AB, y las otras, dos arcos circulares congruentes. El resultado aparece en la figura 251. Si los planos de los arcos forman un ángulo menor de 120°, se obtienen tres superficies que se cortan bajo ángulos de 120°; si se giran los dos arcos, de tal modo que aumente el ángulo formado, la solución cambia continuamente hasta convertirse en dos segmentos circulares planos.

Unamos ahora A y B mediante tres líneas más complicadas; como ejemplo tomaremos tres quebradas, cada una de ellas compuesta por las tres aristas del mismo cubo que unen dos vértices opuestos diagonalmente. Se obtienen así tres superficies congruentes que se cortan

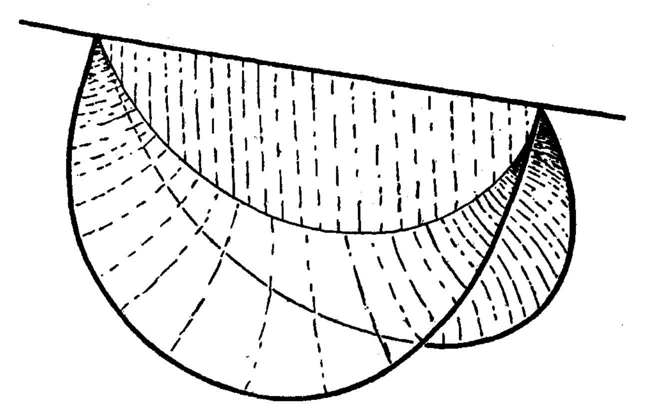


FIG. 251.—Superficies que se cortan bajo ángulos de 120°, tendidas entre tres alambres que unen dos puntos.

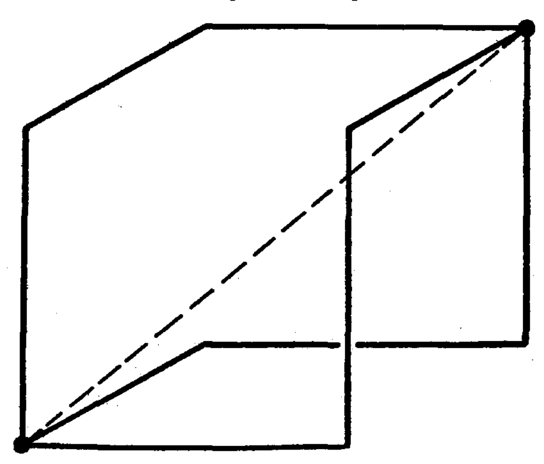


FIG. 252.—Tres quebradas que unen dos puntos.

en la diagonal del cubo. (Se consigue este sistema de superficies a partir del representado en la figura 240, destruyendo las películas adyacentes a tres aristas adecuadamente elegidas.) Si podemos desplazar las quebradas que unen A con B, veremos que la línea de triple intersección se ha convertido en una curva, pero conservándose los ángulos de 120° (Fig. 252).

Todos los casos en que se cortan tres superficies mínimas según curvas determinadas son, fundamentalmente, de naturaleza análoga, y constituyen generalizaciones del problema plano de unir n puntos mediante un sistema mínimo de líneas.

Finalmente, debemos decir algunas palabras sobre las burbujas de jabón. La burbuja esférica de jabón prueba que entre todas las superficies cerradas que incluyen un volumen dado (definido por el aire encerrado en ella), la esfera tiene superficie mínima. Si consideramos burbujas de jabón de volumen dado, que tienden a contraerse hasta adquirir una superficie mínima, pero cuyas transformaciones están restringidas por ciertas condiciones, las superficies resultantes no

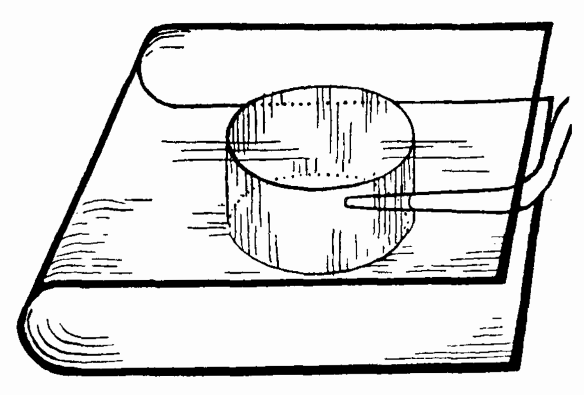


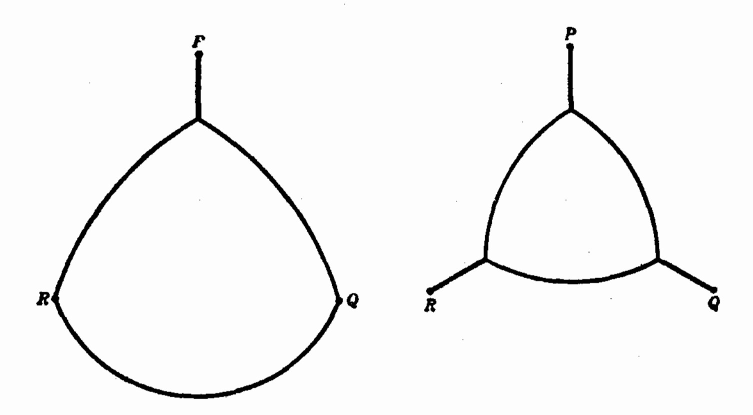
FIG. 253.—Demostración de que el círculo es la figura de perímetro mínimo entre las de igual área.

serán esferas, sino superficies de curvatura media constante, de las cuales la esfera y el cilindro circular constituyen ejemplos particulares.

Por ejemplo, introduzcamos una burbuja de jabón entre dos planos paralelos de cristal, que se han humedecido previamente con la solución jabonosa. Cuando la burbuja toca una de las placas, adopta súbitamente la forma de una semiesfera. En cuanto toca la otra, se transforma instantáneamente en un cilindro circular, demostrando así, de notabilísima manera, la propiedad isoperimétrica del círculo. La clave de este experimento consiste en que la película de solución jabonosa se ajusta verticalmente a la superficie. Formando burbujas de jabón entre dos placas unidas entre sí mediante barras perpendiculares, se puede dar una demostración experimental de los problemas discutidos en las páginas 387-89.

Podemos estudiar el comportamiento de la solución del problema isoperimétrico aumentando o disminuyendo el contenido de aire de

la burbuja mediante un tubo provisto de una punta muy fina. Absorbiendo el aire, no obtenemos, sin embargo, las figuras de la página 388, que consisten en arcos circulares tangentes entre sí. Al disminuir el volumen de aire contenido, los ángulos del triángulo curvilíneo no disminuirán (teóricamente) por debajo de 120º. Resultan las configuraciones indicadas en las figuras 254 y 255, que, a su vez, tienden a los segmentos rectilíneos de la figura 235 al tender el área a cero. La razón matemática de la imposibilidad de formar arcos tangentes con soluciones jabonosas radica en el hecho de que tan pronto como



FIGS. 254 y 255.—Figuras isoperimétricas con condiciones de contorno.

la burbuja se separa de los vértices, las líneas que los unen no deben contarse dos veces. Las figuras 256 y 257 ilustran el resultado de estos experimentos.

***Ejercicio:** Estúdiese el correspondiente problema matemático de hallar un triángulo curvilíneo de área dada y tal que su perímetro, más los tres segmentos que unen los vértices a tres puntos dados, tenga longitud mínima.

Una estructura cúbica, en cuyo interior insuflemos una burbuja, proporcionará superficies de curvatura media constante y de base cuadrática, si la burbuja se expande fuera de la estructura. Si se extrae aire de ella mediante una pajita, se obtienen bellísimas estructuras que se transforman en la representada en la figura 258. Los fenómenos de estabilidad y transición entre los diferentes estados de equilibrio son una fuente de experimentos sumamente instructivos desde el punto de vista matemático. Los experimentos aclaran la teoría de los valores estacionarios, puesto que puede conseguirse que

las transiciones sean de tal naturaleza que conduzcan a un equilibrio inestable, lo que constituye un *estado estacionario.*

Por ejemplo, en la configuración cúbica de la figura 240 aparece una simetría, ya que un plano vertical en el centro une las doce superficies que salen de las aristas. De ahí que deban existir por lo menos otras dos posiciones de equilibrio, una con un cuadrado horizontal y la

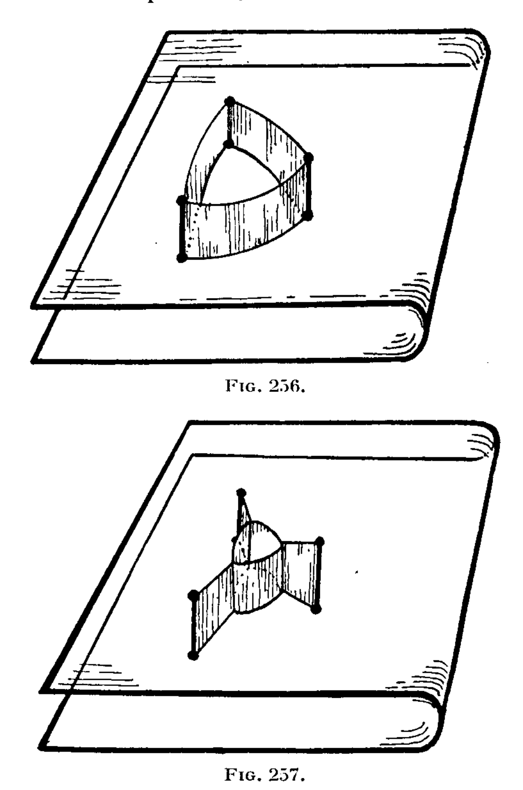


Fig. 256.

Fig. 257.

otra con un cuadrado vertical. En efecto: soplando aire contra las aristas de este cuadrado mediante un tubo muy fino, se puede reducir la estructura a una posición tal que el cuadrado se reduzca a un punto: el centro del cubo. Esta posición de equilibrio inestable se transforma inmediatamente en otra estable, que se deduce de la primera mediante una rotación de 90°.

Puede efectuarse un experimento similar que demuestra el resultado ya obtenido del problema de Steiner para cuatro puntos que forman un cuadrado (Figs. 219 y 220).

Si deseamos obtener las soluciones de tales problemas como casos límites de los isoperimétricos (p. ej., si queremos conseguir la figura 240 a partir de la 258), debemos extraer parte del aire contenido en la burbuja. Pero la figura 258 es completamente simétrica, y su límite, al tender a cero el contenido de la burbuja, será un sistema simétrico de 12 planos, que se cortan en el centro. Es posible observar

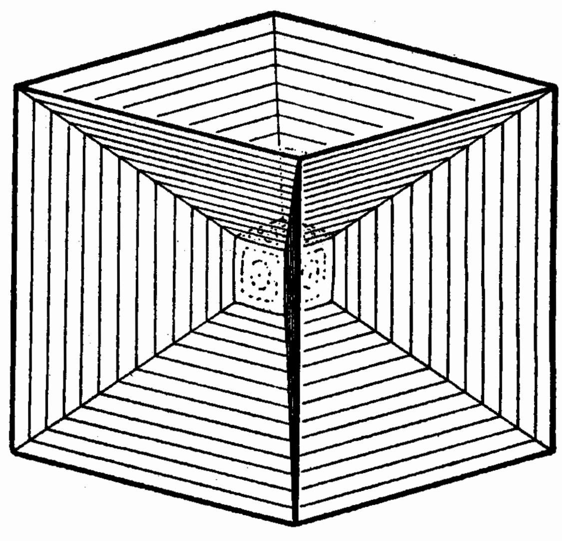


FIG. 258.

realmente este fenómeno, pero la configuración que se obtiene como límite no es estable y se transformará en una de las representadas en la figura 240. Utilizando un líquido algo más viscoso que el descrito anteriormente, puede observarse fácilmente todo el fenómeno. Lo expuesto demuestra que ni siquiera en el terreno de la física la solución de un problema precisa depender continuamente de los datos iniciales; pues en el caso límite correspondiente al volumen cero, la solución de la figura 240 no es límite de la dada por la 258, para el volumen ε, cuando ε tiende a cero.

CAPÍTULO VIII

EL CÁLCULO INFINITESIMAL

Introducción.—Con una simplificación absurda y excesiva de los hechos, se atribuye a veces a dos hombres, Newton y Leibniz, la «invención» del cálculo infinitesimal. En realidad, es el producto de una larga evolución que ni iniciaron ni dieron fin, pero en la cual desempeñaron un papel decisivo. Diseminado por la Europa del siglo XVII se encontraba un grupo de científicos animosos, que en su mayor parte no pertenecían a ningún centro de enseñanza, y que se proponían continuar la obra matemática de Galileo y Kepler. Estos hombres mantenían estrecho contacto entre sí mediante correspondencia postal y viajes. Su atención se concentraba alrededor de dos problemas principales. Primero, el *problema de las tangentes*: la determinación de las tangentes a una curva dada; esto es, el problema fundamental del cálculo diferencial. Segundo, el *problema de las cuadraturas*: determinar el área encerrada por una curva dada; o sea, el problema fundamental del cálculo integral. El gran mérito de Newton y Leibniz consiste en haber reconocido claramente la íntima *conexión entre ambos problemas*. En sus manos, los nuevos métodos unificados se convirtieron en poderosos instrumentos científicos. Gran parte del éxito se debió a la maravillosa notación simbólica ideada por Leibniz. Su éxito no queda disminuído en forma alguna por el hecho de hallarse entremezclada su obra con ideas poco claras e insostenibles, que pudieron inducir a perpetuar la falta de precisión en aquellas mentes que prefieren el misticismo a la claridad. Newton, que indudablemente era de superior categoría científica, parece haber experimentado una intensa influencia de Barrow (1630-1677), su maestro y predecesor en Cambridge. Leibniz era más bien un *dilettante*. Abogado brillante, diplomático, filósofo, y una de las inteligencias más activas y versátiles de su siglo, aprendió la nueva matemática, en un tiempo increíblemente corto, del físico Huygens, mientras se encontraba en París en una misión diplomática. Muy poco tiempo después publicó resultados que contienen el núcleo del cálculo infinitesimal moderno. Newton, cuyos descubrimientos son muy anteriores, se resistió a publicarlos. Además, aunque había encontrado muchos de los resultados en su obra maestra, los *Principia*, por los métodos del cálculo, prefirió presentarlos en el estilo de la geometría clásica, por lo que casi no aparece rastro del

cálculo infinitesimal, al menos en forma explícita, en los *Principia*. Sólo más tarde se publicaron sus investigaciones sobre el método de «fluxiones». Muy pronto sus admiradores iniciaron una guerra cruel sobre cuestiones de prioridad con los amigos de Leibniz. Acusaron a éste último de plagiario, aunque en una atmósfera saturada de las nuevas teorías nada era más natural que el descubrimiento simultáneo e independiente. Aquella polémica sobre la prioridad en la «invención» del cálculo constituye un desdichado ejemplo del valor excesivo dado a las cuestiones de precedencia y derecho de propiedad intelectual, muy adecuadas para envenenar la atmósfera de la vida científica.

En el análisis matemático del siglo XVII y gran parte del XVIII pareció haberse abandonado por completo el ideal griego de razonamiento claro y riguroso. La «intuición» y el «instinto» reemplazaron a la razón en numerosos casos importantes. Esto fomentó la fe sin discriminación en el poder sobrehumano de los nuevos métodos. Se creía generalmente no sólo innecesaria, sino imposible, una clara presentación de los resultados del cálculo infinitesimal. De no haber estado la nueva ciencia en manos de un grupo restringido de hombres extremadamente competentes, habrían resultado serios errores y aun desastres. Estos científicos se dejaban guiar por un sentido profundamente instintivo, que les impedía perderse por completo. Pero cuando la Revolución francesa abrió el camino hacia una inmensa extensión de la cultura superior, cuando aumentó el número de personas que deseaban participar en la actividad científica, ya no pudo posponerse la revisión crítica del nuevo análisis. El siglo XIX llevó felizmente a cabo esa tarea, y hoy día puede enseñarse el cálculo infinitesimal sin ninguna traza de misterio y con completo rigor. No existe actualmente ninguna razón para que las personas educadas no entiendan este instrumento básico de las ciencias.

En este capítulo nos proponemos dar una introducción elemental, en la cual insistiremos en la adecuada comprensión de los conceptos fundamentales, más que en la manipulación formal. Utilizaremos constantemente un lenguaje intuitivo, pero siempre de manera que se mantenga la conexión con los conceptos precisos y la claridad del procedimiento.

I. LA INTEGRAL

1. El área como límite.—Para poder calcular el área de una figura plana elegimos como *unidad de área* un cuadrado de lado igual a la unidad de longitud. Si ésta es un centímetro, la unidad correspondiente de área será el centímetro cuadrado; es decir, un cuadrado de

lado igual a un centímetro. Partiendo de esta definición es muy fácil calcular el área de un rectángulo. Si p y q son las longitudes de dos lados adyacentes, medidas con la unidad de longitud, la superficie del rectángulo es igual a pq unidades de área, o más brevemente, el área es igual al producto pq. Esto es válido para p y q arbitrarios, racionales o no. Si p y q son racionales, se obtiene este resultado escribiendo $p = m/n$, $q = m'/n'$, siendo m, n, m' y n' números enteros. Hallaremos entonces la medida común $1/N = 1/nn'$ de ambos lados, por lo que $p = mn' \cdot 1/N$, $q = nm' \cdot 1/N$. Finalmente, subdividiremos el rectángulo en pequeños cuadrados de lado $1/N$ y área $1/N^2$. El número de tales cuadrados será $nm' \cdot mn'$, y el área total, $nm'mn' \cdot 1/N^2 = nm'mn'/n^2n'^2 = m/n \cdot m'/n' = pq$. Si p y q son irracionales, se obtiene el mismo resultado reemplazando primero p y q por números racionales aproximados p_r y q_r, respectivamente, y haciendo después tender p_r y q_r hacia p y q.

Geométricamente es evidente que el área de un triángulo es igual a la mitad de la del rectángulo con la misma base, b, y la misma altura, h, por lo que el área del triángulo viene dada por la expresión usual $^1/_2\, bh$. Cualquier dominio plano, limitado por una o varias poligonales, puede descomponerse en triángulos, por lo que su área puede obtenerse como suma de las de éstos.

Se plantea la necesidad de un método más general de calcular áreas cuando investigamos el área de una figura no limitada por rectas, sino por *curvas*; p. ej., ¿cómo determinaremos el área de un disco circular o de un segmento de parábola? Esta cuestión crucial, punto de partida del cálculo integral, fué objeto de las investigaciones de Arquímedes, en el siglo III a. de J. C., quien las calculó mediante un proceso «exhaustivo». Con Arquímedes y los grandes matemáticos hasta la época de Gauss, podemos adoptar la «ingenua» actitud de que las áreas curvilíneas son entes dados intuitivamente, y que la cuestión no consiste en *definirlas*, sino en *calcularlas* (véase, sin embargo, la discusión de la pág. 473). Inscribamos en el dominio otro dominio aproximado de contorno poligonal, cuya área está bien definida, y que representa aproximadamente al primero. Eligiendo otro dominio poligonal que incluya al anterior, obtenemos una aproximación mejor del dominio dado, y procediendo de esta manera podemos paulatinamente *agotar* toda la superficie y obtener el área buscada como límite de las áreas de una sucesión de dominios poligonales inscritos cuyo número de lados crece indefinidamente. De esta manera puede calcularse el área del círculo de radio 1; su valor numérico se designa mediante el símbolo π.

Arquímedes siguió este sistema para obtener el área del círculo y del segmento parabólico. Durante el siglo XVII se resolvieron así muchos otros casos. En cada uno de ellos el cálculo efectivo del límite dependía de algún procedimiento ingenioso, especialmente adaptado para ese problema particular. Uno de los principales éxitos del cálculo consistió en reemplazar esos procedimientos especiales y restringidos de obtener el área por un método potente y general.

2. La integral.—El primer concepto fundamental del cálculo es el de integral. En lo que sigue entenderemos la integral como la expresión mediante un límite del *área limitada por una curva.* Dada una función continua y positiva, $y = f(x)$; p. ej., $y = x^2$ o $y = 1 + \cos x$, consideremos el dominio así limitado: inferiormente, por el segmento del eje de las x, desde la abscisa a hasta otra mayor b; lateralmente, por las perpendiculares al eje de las x en esos puntos; y superiormente, por la curva $y = f(x)$. Nos proponemos calcular el área A de ese dominio.

Puesto que, en general, ese dominio no puede descomponerse en rectángulos o triángulos, no se dispone de una expresión inmediata del área A que permita su cálculo explícito. Pero podemos calcular un valor aproximado de A, y representar así A como un límite, de la siguiente manera: dividimos el intervalo desde $x = a$ hasta $x = b$ en cierto número de pequeños subintervalos; levantamos perpendiculares en cada uno de los puntos de subdivisión, y reemplazamos cada franja del dominio bajo la curva por un rectángulo cuya altura se elige entre las ordenadas máxima y mínima de la curva en esa franja. La suma S de las áreas de estos rectángulos proporciona un valor aproximado de la verdadera área A bajo la curva. La exactitud de este procedimiento será tanto mejor cuanto mayor sea el número de rectángulos y más pequeña la anchura de cada uno de ellos. Así podemos definir el área verdadera como un límite; si formamos una sucesión,

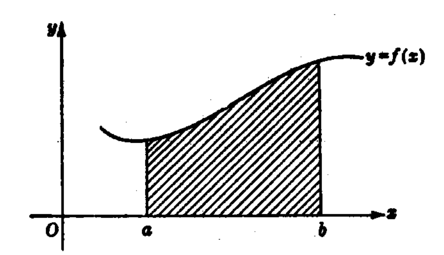


FIG. 259.—La integral como área.

$$S_1, S_2, S_3, \cdots, \qquad [1]$$

de aproximaciones rectangulares al área bajo la curva, de tal manera que la base del rectángulo más ancho de S_n tienda a 0 al aumentar n, entonces la sucesión [1] tiende al límite A; es decir:

$$S_n \to A, \qquad [2]$$

siendo este límite A, o sea, el área bajo la curva, independiente de la elección particular de la sucesión [1], siempre que las anchuras de los rectángulos de aproximación tiendan a cero. (Así, p. ej., puede deducirse S_n de S_{n-1} agregando uno o más puntos nuevos de subdivisión a los que definen S_{n-1}, o bien la elección de los puntos de subdivisión de S_n puede ser enteramente independiente de la correspondiente a S_{n-1}.) Por definición, llamamos *integral de la función* f(x) *desde* a *hasta* b al área A del dominio expresada por este proceso de límite, y se designa por un símbolo especial, el «signo integral», que se escribe:

$$A = \int_a^b f(x)\,dx. \qquad [3]$$

Leibniz introdujo el símbolo $\int$, la «dx» y el nombre de «integral» para sugerir la manera como se obtiene el límite. Para explicar esta notación debemos repetir con más detalle el proceso de aproximación al área A. Al propio tiempo, la formulación analítica del proceso de límite nos permitirá prescindir de las hipótesis restrictivas $f(x) \geqslant 0$ y $b > a$, y eliminar finalmente el concepto intuitivo de área como base de nuestra definición de integral (esto último se hará en el suplemento a este capítulo).

Dividamos el intervalo (a, b) en n intervalos parciales, que, para mayor sencillez, supondremos todos de igual amplitud, $(b - a)/n$. Designaremos los puntos de subdivisión por

$$x_0 = a, \qquad x_1 = a + \frac{b-a}{n},$$

$$x_2 = a + \frac{2(b-a)}{n}, \cdots, x_n = a + \frac{n(b-a)}{n} = b.$$

Para representar la cantidad $(b - a)/n$, diferencia entre dos valores de x consecutivos, introduciremos la notación Δx (léase «delta x»):

$$\Delta x = \frac{b-a}{n} = x_{j+1} - x_j,$$

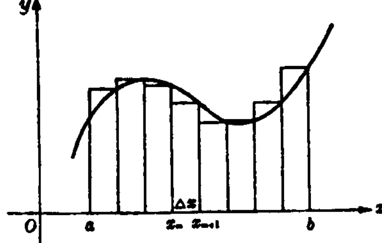


FIG. 260.—Aproximación del área por pequeños rectángulos.

en la cual el símbolo Δ significa simplemente «diferencia» (es un «operador» y no debe confundirse con un número). Podemos elegir como altura de cada rectángulo de aproximación el valor $y = f(x)$ en el extremo derecho del subintervalo; entonces la suma de las áreas de estos rectángulos será:

$$S_n = f(x_1) \cdot \Delta x + f(x_2) \cdot \Delta x + \cdots + f(x_n) \cdot \Delta x, \qquad [4]$$

que se escribe abreviadamente así:

$$S_n = \sum_{j=1}^{n} f(x_j) \cdot \Delta x. \qquad [5]$$

Aquí el símbolo $\sum_{j=1}^{n}$ (léase «sigma desde $j = 1$ hasta n») representa la suma de todas las expresiones obtenidas al dar a j sucesivamente los valores 1, 2, 3, 4, ..., n.

Puede verse la utilidad del uso del símbolo $\sum$ para expresar en forma concisa el resultado de una suma en los ejemplos siguientes.

$$2 + 3 + 4 + \cdots + 10 = \sum_{j=2}^{10} j,$$

$$1 + 2 + 3 + \cdots + n = \sum_{j=1}^{n} j,$$

$$1^2 + 2^2 + 3^2 + \cdots + n^2 = \sum_{j=1}^{n} j^2,$$

$$aq + aq^2 + \cdots + aq^n = \sum_{j=1}^{n} aq^j,$$

$$a + (a + d) + (a + 2d) + \cdots + (a + nd) = \sum_{j=0}^{n} (a + jd).$$

Formemos ahora una sucesión de dichas aproximaciones S_n en las cuales n aumenta indefinidamente, con lo que el número de términos en cada suma [5] aumenta, mientras que cada uno de los términos $f(x_j)\Delta x$ tiende a cero, debido a la presencia del factor $\Delta x = (b - a)/n$. Al aumentar n esta suma tiende al área A,

$$A = \lim \sum_{j=1}^{n} f(x_j)\Delta x = \int_a^b f(x)\, dx. \qquad [6]$$

Leibniz expresó simbólicamente este paso al límite de la suma aproximada S_n a A reemplazando el signo de sumación $\sum$ por $\int$, y el símbolo de diferencia Δ por d. (El símbolo $\sum$ de sumación se escribía generalmente S en tiempos de Leibniz, y el símbolo $\int$ es simplemente una S estilizada.) Aunque el simbolismo de Leibniz sugiere muy

bien la manera de obtener la integral como límite de una suma finita es necesario tener cuidado y no dar demasiada importancia a lo que, después de todo, es sólo un puro convenio respecto a la forma de expresión del límite. En los primeros tiempos del cálculo, cuando todavía no se había entendido de una forma precisa el concepto de límite y ciertamente no siempre se le tenía presente, se solía explicar el concepto de integral diciendo que «se reemplaza la diferencia finita Δx por la cantidad infinitamente pequeña dx, siendo la integral la suma de las infinitas cantidades infinitamente pequeñas $f(x)\, dx$». Aunque lo infinitamente pequeño tiene un cierto atractivo para los espíritus inclinados a la especulación, este concepto ha sido desplazado de la matemática moderna. No se llega a ningún resultado útil rodeando la noción clara de integral de una niebla de frases sin sentido. El mismo Leibniz se dejó arrastrar muchas veces por el poder de sugestión de sus símbolos, que actúan *como si* denotaran una suma de cantidades «infinitamente pequeñas», con las cuales, sin embargo, se puede operar hasta cierto punto como con las cantidades ordinarias. En efecto, se acuñó el vocablo integral para indicar que la totalidad del área A se compone de partes «infinitesimales» $f(x)\, dx$. De todas formas, hubo de pasar casi un siglo, después de Leibniz y Newton, para comprender claramente que la definición de integral se basa sólo en el concepto de límite. Permaneciendo firmemente en dicha base podemos evitar todas las nebulosidades, todas las dificultades y todos los disparates que tanto perturbaron el desarrollo del cálculo en sus comienzos.

3. Observaciones generales sobre el concepto de integral. Definición general.—En nuestra definición geométrica de la integral como área hemos supuesto explícitamente que $f(x)$ no es negativa en el intervalo $[a, b]$ de integración; es decir, que ninguna parte de la gráfica de la curva se encuentra por debajo del eje de las x. Pero, en nuestra definición analítica de la integral como límite de una sucesión de sumas S_n, esa hipótesis resulta superflua. Tomamos simplemente las pequeñas cantidades $f(x_j) \cdot \Delta x$, formamos su suma y pasamos al límite, procedimiento que sigue teniendo sentido aunque algunos de los valores $f(x_j)$ sean negativos. La interpretación geométrica mediante el área (Fig. 261) nos dice que la integral de $f(x)$ es la suma *algebraica* de las áreas limitadas

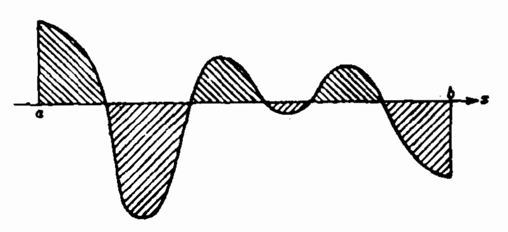


FIG. 261.—Áreas positivas y negativas.

por la gráfica y el eje de las x, contando las que se encuentran por encima del eje como positivas, y como negativas las que se encuentran por debajo.

Puede ocurrir que en las aplicaciones aparezcan integrales $\int_a^b f(x)\,dx$ en las que b sea menor que a, de forma que la diferencia $(b - a)/n = \Delta x$ resulte negativa. En nuestra definición analítica tenemos que $f(x_j)\cdot\Delta x$ es negativo si $f(x_j)$ es positiva y Δx negativo, etc. En otras palabras, el valor de la integral será el valor opuesto del de la integral de b a a. Así, pues, se tiene la sencilla regla:

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx.$$

Debemos subrayar que el valor de la integral sigue siendo el mismo, aunque nos limitemos a puntos de subdivisión equidistantes; o lo que es lo mismo, a diferencias de x iguales $\Delta x = x_{j+1} - x_j$. Podemos elegir los x_j de otro modo, de forma que las diferencias $\Delta x_j = x_{j+1} - x_j$ no sean iguales entre sí (por lo que será necesario distinguirlas mediante subíndices). Aun en este caso las sumas

$$S_n = f(x_1)\Delta x_0 + f(x_2)\Delta x_1 + \cdots + f(x_n)\Delta x_{n-1}$$

y también las sumas

$$S'_n = f(x_0)\Delta x_0 + f(x_1)\Delta x_1 + \cdots + f(x_{n-1})\Delta x_{n-1}$$

tienden al mismo límite; o sea, al valor de la integral $\int_a^b f(x)dx$, bastando tomar la precaución de que, al aumentar n, todas las diferencias $\Delta x_j = x_{j+1} - x_j$ tiendan a cero, de tal manera que la mayor de ellas para cierto valor de n tienda a cero al aumentar n.

De acuerdo con todo esto, la *definición definitiva de integral* está dada por

$$\int_a^b f(x)\,dx = \lim \sum_{j=1}^{n} f(v_j)\,\Delta x_j \qquad [6a]$$

cuando $n \to \infty$. En este límite, v_j puede indicar cualquier punto del intervalo $x_j \leqslant v_j \leqslant x_{j+1}$, con la única restricción de que el mayor intervalo $\Delta x_j = x_{j+1} - x_j$ tienda a cero al crecer n.

La existencia del límite [6a] no necesita demostración si suponemos definido el concepto de área bajo una curva y admitimos la posibilidad de aproximarse a ella mediante sumas de rectángulos. Sin embargo, como se deducirá más adelante (pág. 473), un análisis más atento pondría de manifiesto que es deseable e incluso necesario para presentar de una manera lógica y completa la noción de integral de-

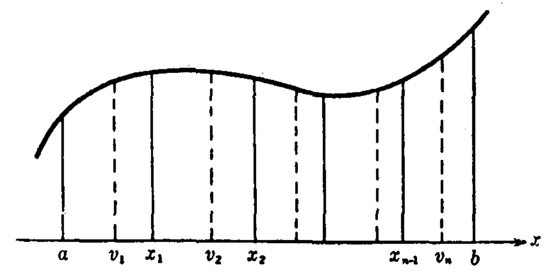


FIG. 262.—Subdivisión arbitraria en la definición general de integral.

mostrar la existencia del límite para cualquier función continua $f(x)$, sin referencia al concepto geométrico de área.

4. Ejemplos de integración. Integración de x^r.—Hasta ahora, nuestra discusión de la integral ha sido meramente teórica. La cuestión principal consiste en saber si formando una suma S_n y pasando al límite se llega realmente a resultados tangibles en casos concretos. Naturalmente, esto requerirá algún razonamiento adicional adaptado a la función particular $f(x)$ cuya integral se busca. Cuando Arquímedes, hace 2000 años, encontró el área del segmento parabólico, realizó lo que actualmente llamamos la integración de la función $f(x) = x^2$, mediante un procedimiento muy ingenioso; en el siglo XVII, los precursores del cálculo integral moderno consiguieron resolver problemas de integración de funciones muy sencillas, tales como x^n, también por medio de procedimientos especiales. Sólo después de adquirir mucha experiencia con numerosos casos especiales se halló en los métodos sistemáticos del cálculo un procedimiento general de tratar el problema de la integración, con lo cual se amplió el campo de los problemas particulares solubles. En los párrafos siguientes discutiremos algunos problemas especialmente instructivos, que pertenecen a la etapa primitiva del cálculo, pues nada será más adecuado para aclarar el paso al límite, que constituye la esencia de la integración.

a) Empecemos por un ejemplo casi trivial. Si $y = f(x)$ es una constante, p. ej., $f(x) = 2$, evidentemente la integral $\int_a^b 2dx$, interpretada como un área, es igual a $2(b - a)$, puesto que el área de un rectángulo es igual a la base por la altura. Comparemos este resultado con la definición de integral como límite, según [6]. Si sustituímos en [5] $f(x_j)$ por 2 para todos los valores de j, encontraremos que

$$S_n = \sum_{j=1}^{n} f(x_j)\Delta x = \sum_{j=1}^{n} 2\Delta x = 2\sum_{j=1}^{n} \Delta x = 2(b - a)$$

para todo n, puesto que

$$\sum_{j=1}^{n} \Delta x = (x_1 - x_0) + (x_2 - x_1) + \cdots + (x_n - x_{n-1}) = x_n - x_0 = b - a.$$

b) Casi tan sencilla resulta la integración de $f(x) = x$. En este caso, $\int_a^b x dx$ es el área de un trapecio (Fig. 263), y ésta, por geometría elemental, vale

$$(b - a)\frac{b + a}{2} = \frac{b^2 - a^2}{2}$$

Este resultado se halla nuevamente de acuerdo con la definición [6] de la integral, como se ve efectuando el paso al límite sin recurrir a la figura geométrica. Si en [5] hacemos $f(x) = x$, la suma S_n se transforma en

$$\begin{aligned} S_n &= \sum_{j=1}^{n} x_j\Delta x = \sum_{j=1}^{n} (a + j\Delta x)\Delta x \\ &= (na + \Delta x + 2\Delta x + 3\Delta x + \cdots + n\Delta x)\Delta x \\ &= na\Delta x + (\Delta x)^2(1 + 2 + 3 + \cdots + n). \end{aligned}$$

Utilizando la fórmula [1] de la página 19, que da la suma de la progresión aritmética $1 + 2 + 3 + ... + n$, obtenemos:

$$S_n = na\Delta x + \frac{n(n + 1)}{2}(\Delta x)^2.$$

Por ser $\Delta x = \dfrac{b - a}{n}$, ésta es igual a

$$S_n = a(b - a) + \tfrac{1}{2}(b - a)^2 + \frac{1}{2n}(b - a)^2.$$

Si ahora hacemos tender n a infinito, el último sumando tiende a cero, y se obtiene:

$$\lim S_n = \int_a^b x\, dx = a(b - a) + {}^1/_2(b - a)^2 = {}^1/_2(b^2 - a^2).$$

de acuerdo con la interpretación geométrica de la integral como área.

c) Menos trivial resulta la integración de la función $f(x) = x^2$. Arquímedes hizo uso de métodos geométricos para resolver el problema equivalente de hallar el área de un segmento de la parábola $y = x^2$. Nosotros vamos a proceder analíticamente, basándonos en la

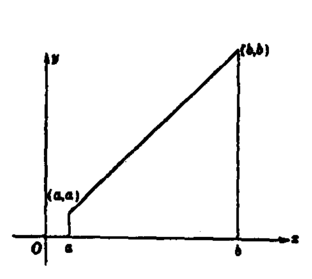


FIG. 263.—Área de un trapecio.

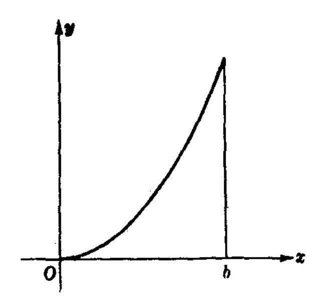


FIG. 264.—Área de un segmento parabólico.

definición [6a]. Para simplificar el cálculo formal elegimos 0 como «límite inferior» a de la integral; entonces $\Delta x = b/n$. Como $x_j = j \cdot \Delta x$ y $f(x_j) = j^2(\Delta x)^2$, obtenemos para S_n la expresión

$$S_n = \sum_{j=1}^{n} f(j\Delta x)\Delta x = [1^2 \cdot (\Delta x)^2 + 2^2 \cdot (\Delta x)^2 + \cdots + n^2(\Delta x)^2] \cdot \Delta x$$
$$= (1^2 + 2^2 + \cdots + n^2)(\Delta x)^3.$$

Ahora podemos calcular efectivamente el límite; utilizando la fórmula

$$1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

establecida en la página 21, y haciendo la sustitución $\Delta x = b/n$, tenemos:

$$S_n = \frac{n(n+1)(2n+1)}{6} \cdot \frac{b^3}{n^3} = \frac{b^3}{6}\left(1 + \frac{1}{n}\right)\left(2 + \frac{1}{n}\right)$$

Esta transformación preliminar facilita el paso al límite, puesto que $1/n$ tiende a cero al crecer n indefinidamente. Así, obtenemos el límite $\frac{b^3}{6} \cdot 1 \cdot 2 = \frac{b^3}{3}$ y resulta, por tanto,

$$\int_0^b x^2\,dx = b^3/3.$$

Aplicando este resultado al área desde 0 a a, tenemos:

$$\int_0^a x^2\,dx = a^3/3,$$

y por sustracción:

$$\int_a^b x^2\,dx = \frac{b^3 - a^3}{3}$$

Ejercicio: Demuéstrese de la misma manera, empleando la fórmula [5] de la página 22, que

$$\int_a^b x^3\,dx = \frac{b^4 - a^4}{4}$$

Utilizando la fórmula general que expresa la suma $1^k + 2^k + \cdots + n^k$ de las potencias k-ésimas de todos los números naturales desde 1 hasta n, se puede obtener el resultado

$$\int_a^b x^k\,dx = \frac{b^{k+1} - a^{k+1}}{k+1}, \text{ siendo } k \text{ un entero positivo arbitrario.} \qquad [7]$$

*En lugar de proceder de esta manera, podemos obtener de forma más sencilla un resultado aún más general utilizando nuestra observación anterior respecto a la posibilidad de calcular la integral mediante una subdivisión del intervalo por puntos no equidistantes. Estableceremos la fórmula [7] no sólo para cualquier entero positivo k, sino para un número racional arbitrario, positivo o negativo,

$$k = u/v,$$

donde u es un entero positivo y v otro entero, positivo o negativo. Se excluirá únicamente el valor $k = -1$, para el cual la fórmula [7] carece de significado. Supondremos, además, que $0 < a < b$.

Para obtener la fórmula integral [7] formamos S_n eligiendo los puntos de subdivisión $x_0 = a, x_1, x_2, \cdots, x_n = b$ en *progresión geométrica*. Pongamos $\sqrt[n]{b/a} = q$, por lo que $b/a = q^n$, y definamos $x_0 = a$, $x_1 = aq$, $x_2 = aq^2, \cdots, x_n = aq^n = b$. Mediante este artificio, el paso al límite se hace muy sencillo. Para la «suma rectangular» S_n encontramos:

$$S_n = a^k(aq - a) + a^k q^k(aq^2 - aq) + a^k q^{2k}(aq^3 - aq^2) + \\ + \cdots + a^k q^{(n-1)k}(aq^n - aq^{n-1}),$$

puesto que $f(x_j) = x_j^k = a^k q^{jk}$ y $\Delta x_j = x_{j+1} - x_j = aq^{j+1} - aq^j$.

Como cada término contiene el factor $a^k(aq - a)$, podemos escribir:

$$S_n = a^{k+1}(q - 1)\{1 + q^{k+1} + q^{2(k+1)} + \cdots + q^{(n-1)(k+1)}\}.$$

Sustituyendo q^{k+1} por t vemos que la expresión entre corchetes es una progresión geométrica, $1 + t + t^2 + \cdots + t^{n-1}$, cuya suma, como se demostró en la página 21, es $(t^n - 1)/(t - 1)$. Pero $t^n = q^{n(k+1)} = \left(\frac{b}{a}\right)^{k+1} = \frac{b^{k+1}}{a^{k+1}}$.
En consecuencia,

$$S_n = (q - 1)\left\{\frac{b^{k+1} - a^{k+1}}{q^{k+1} - 1}\right\} = \frac{b^{k+1} - a^{k+1}}{N} \qquad [8]$$

donde

$$N = \frac{q^{k+1} - 1}{q - 1}$$

Hasta aquí n ha permanecido constante. Hagamos crecer n y determinemos el límite de N. Al aumentar n, la raíz n-ésima $\sqrt[n]{b/a} = q$ tenderá a 1 (pág. 334), por lo que tanto el numerador como el denominador de N tenderán a cero, lo que impone cierta cautela. Supongamos primero que k es un entero positivo; podrá efectuarse entonces la división por $q - 1$, obteniendo (pág. 21) $N = q^k + q^{k-1} + \cdots + q + 1$. Si ahora n aumenta, q tiende a 1, o sea, que $q^2, q^3 \cdots, q^k$ tenderán también a 1, por lo que N tiende a $k + 1$. Pero esto nos dice que S_n tiende a $\frac{b^{k+1} - a^{k+1}}{k+1}$, según queríamos demostrar.

Ejercicio: Demuéstrese para cualquier número racional $k \neq -1$ que el límite es el mismo; es decir, que $N \to k + 1$ y que, en consecuencia, el resultado [7] continúa siendo válido. Siguiendo los mismos pasos, demuéstrese primero para números enteros negativos k. Después, si $k = u/v$, escríbase $q^{1/v} = s$, resultando:

$$N = \frac{s^{(k+1)v} - 1}{s^v - 1} = \frac{s^{u+v} - 1}{s^v - 1} = \frac{s^{u+v} - 1}{s - 1} : \frac{s^v - 1}{s - 1}$$

Si n aumenta, tanto s como q tienden a 1, por lo que los dos cocientes del segundo miembro tienden a $u + v$ y v, respectivamente, lo que proporciona nuevamente $\frac{u+v}{v} = k + 1$ como límite de N.

Más adelante veremos cómo los métodos más simples y potentes del cálculo reemplazan esta discusión un tanto larga y artificiosa.

Ejercicios:

1. Verifíquese la integración precedente de x^k para los casos $k = \frac{1}{2}, -\frac{1}{2}$, 2, -2, 3, -3.

2. Calcúlese el valor de las siguientes integrales:

$$\text{a)} \int_{-2}^{-1} x\,dx; \quad \text{b)} \int_{-1}^{+1} x\,dx; \quad \text{c)} \int_{1}^{2} x^2\,dx; \quad \text{d)} \int_{-1}^{-2} x^3\,dx; \quad \text{e)} \int_{0}^{n} x\,dx.$$

3. Determínese el valor de las siguientes integrales:

$$\text{a)} \int_{-1}^{+1} x^3\,dx; \quad \text{b)} \int_{-2}^{2} x^3 \cos x\,dx; \quad \text{c)} \int_{-1}^{+1} x^4 \cos^2 x \operatorname{sen}^5 x\,dx; \quad \text{d)} \int_{-1}^{+1} \operatorname{tg} x\,dx.$$

(Indicación: Considérese la gráfica de cada una de las funciones bajo el signo integral; téngase en cuenta su simetría con respecto a $x = 0$, e interprétense las integrales como áreas.)

*4. Intégrese sen x y cos x desde 0 hasta b, haciendo $\Delta x = h$ y utilizando las fórmulas dadas en la página 498.

5. Intégrese $f(x) = x$ y $f(x) = x^2$ desde 0 hasta b, dividiendo el intervalo en partes iguales y utilizando en [6a] los valores $v_j = {}^1/_2(x_j + x_{j+1})$.

*6. Utilizando el resultado [7] y la definición de integral con valores iguales de Δx, demuéstrese para $n \to \infty$ la relación límite:

$$\frac{1^k + 2^k + \cdots + n^k}{n^{k+1}} \to \frac{1}{k+1} \text{ para } n \to \infty.$$

(Hágase $\frac{1}{n} = \Delta x$ y demuéstrese que el límite es igual a $\int_0^1 x^k\,dx$.)

7. Demuéstrese, para $n \to \infty$, que

$$\frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{1+n}} + \frac{1}{\sqrt{2+n}} + \cdots + \frac{1}{\sqrt{n+n}}\right) \to 2(\sqrt{2} - 1).$$

(Escríbase esta suma de tal modo que el límite aparezca como una integral.)

8. Calcúlese el área del segmento parabólico limitado por un arco $P P_2$ y la cuerda correspondiente de una parábola $y = ax^2$ en función de las abscisas x_1 y x_2 de los dos puntos.

5. Reglas del «cálculo integral».—Un paso importante en el desarrollo del cálculo consistió en formular ciertas reglas generales mediante las cuales podían reducirse problemas complicados a otros más sencillos, con lo que la solución aparecía de una manera casi mecánica. Este carácter algorítmico se pone especialmente de manifiesto con la notación de Leibniz. Sin embargo, una atención excesiva a la parte puramente mecánica de la resolución de los problemas puede degradar la enseñanza del cálculo hasta convertirla en una disciplina vacía.

De la definición [6] o de la interpretación geométrica de la integral como área se deducen inmediatamente algunas reglas sencillas de integración.

La integral de la suma de dos funciones es igual a la suma de las integrales de las dos funciones. La integral del producto de una constante c *por una función* f(x) *es igual a* c *por la integral de* f(x). La fórmula siguiente combina ambas reglas:

$$\int_a^b [cf(x) + dg(x)]\, dx = c\int_a^b f(x)\, dx + d\int_a^b g(x)\, dx. \qquad [9]$$

La demostración es consecuencia inmediata de la definición de integral como límite de la suma finita [5], puesto que la fórmula correspondiente para una suma S_n es evidentemente cierta. La regla se extiende en seguida a sumas de más de dos funciones.

Como ejemplo del uso de la misma consideremos un polinomio,

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n,$$

cuyos coeficientes $a_0, a_1, a_2, \ldots, a_n$ son constantes. Para formar la integral de $f(x)$ desde a hasta b, procederemos término a término, de acuerdo con la regla.

Utilizando la fórmula [7], encontramos:

$$\int_a^b f(x)\, dx = a_0(b - a) + a_1\frac{b^2 - a^2}{2} + \cdots + a_n\frac{b^{n+1} - a^{n+1}}{n+1}$$

Otra regla análoga, consecuencia evidente tanto de la definición analítica como de la interpretación geométrica de la integral, es la siguiente:

$$\int_a^b f(x)\,dx + \int_b^c f(x)\,dx = \int_a^c f(x)\,dx. \qquad [10]$$

Además es evidente que la integral es igual a cero si a es igual a b. La regla de la página 415,

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx, \qquad [11]$$

está de acuerdo con las dos últimas enunciadas aquí, puesto que corresponde a [10] para $c = a$.

A veces conviene utilizar el hecho de la independencia de la integral respecto de la letra elegida para designar la variable; p. ej.,

$$\int_a^b f(x)\,dx = \int_a^b f(u)\,du = \int_a^b f(t)\,dt, \text{ etc.},$$

ya que un simple cambio en el nombre de las coordenadas del sistema al cual se refiere la gráfica de la función no altera el área bajo la curva. La misma observación vale si introducimos ciertos cambios en el propio sistema de coordenadas; p. ej., si trasladamos el origen hacia la derecha en una unidad, de O a O′, como en la figura 265, con lo cual x queda reemplazada por una nueva abscisa x' tal que $x = 1 + x'$. Una curva de ecuación $y = f(x)$ tendrá en el nuevo sistema de coordenadas la ecuación $y = f(1 + x')$. [Por ejemplo, $y = 1/x = 1/(1 + x')$]. Si el área A dada bajo esta curva estaba limitada por $x = 1$ y $x = b$, en el nuevo sistema de coordenadas quedará limitada por $x' = 0$ y $x' = b - 1$. Así se tiene:

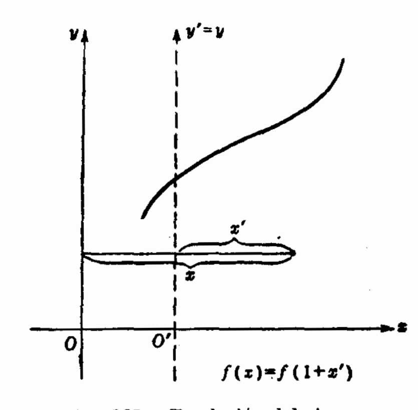


FIG. 265.—Traslación del eje y.

$$\int_1^b f(x)\,dx = \int_0^{b-1} f(1 + x')\,dx',$$

o bien, cambiando x' por u,

$$\int_{1}^{b} f(x)\,dx = \int_{0}^{b-1} f(1+u)\,du; \qquad [12]$$

por ejemplo,

$$\int_{1}^{b} \frac{1}{x}\,dx = \int_{0}^{b-1} \frac{1}{1+u}\,du; \qquad [12a]$$

y para la función $f(x) = x^k$,

$$\int_{1}^{b} x^k\,dx = \int_{0}^{b-1} (1+u)^k\,du. \qquad [12b]$$

Análogamente,

$$\int_{0}^{b} x^k\,dx = \int_{-1}^{b-1} (1+u)^k\,du \qquad (k > 0). \qquad [12c]$$

Puesto que el primer miembro de [12c] es igual a $b^{k+1}/(k+1)$, tendremos:

$$\int_{-1}^{b-1} (1+u)^k\,du = \frac{b^{k+1}}{k+1} \qquad [12d]$$

Ejercicios:

1. Calcúlese la integral del $1 + x + x^2 + \cdots + x^n$ desde 0 hasta b.
2. Demuéstrese que para $n > 0$ la integral de $(1 + x)^n$ desde -1 hasta z es igual a

$$\frac{(1+z)^{n+1}}{(n+1)}$$

3. Demuéstrese que la integral desde 0 hasta 1 de x^n sen x es menor que $1/(n + 1)$. (Este último valor es el de la integral de x^n.)
4. Demuéstrese directamente y utilizando el teorema del binomio, que la integral desde -1 a z de $(1 + x)^n/n$, es $(1 + z)^{n+1}/n\,(n + 1)$.

Finalmente, mencionaremos dos reglas importantes que aparecen en forma de desigualdades y que permiten estimar de forma poco aproximada, pero útil, el valor de muchas integrales.

Supongamos $b > a$ y que los valores de $f(x)$ en el intervalo no sobrepasan a los de otra función $g(x)$. Entonces se tiene:

$$\int_{a}^{b} f(x)\,dx < \int_{a}^{b} g(x)\,dx, \qquad [13]$$

como se deduce inmediatamente de la figura 266 o de la definición analítica de integral. En particular, si $g(x) = M$ es una constante no excedida en ningún punto por los valores de $f(x)$, tendremos:

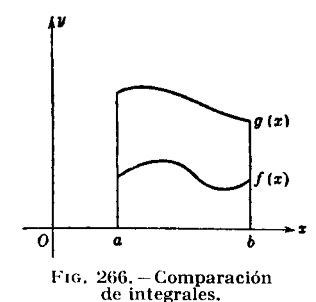


FIG. 266. — Comparación de integrales.

$$\int_a^b g(x)\,dx = \int_a^b M\,dx = M\,(b - a).$$

Por consiguiente,

$$\int_a^b f(x)\,dx \leqslant M\,(b - a). \qquad [14]$$

Si $f(x)$ es no negativa, será $f(x) = |f(x)|$. Si $f(x) < 0$, será $|f(x)| > f(x)$; de donde resulta, poniendo $g(x) = |f(x)|$ en [13], que

$$\int_a^b f(x)\,dx \leqslant \int_a^b |f(x)|\,dx. \qquad [15]$$

Puesto que $|-f(x)| = |f(x)|$, se deduce también que

$$-\int_a^b f(x)\,dx \leqslant \int_a^b |f(x)|\,dx,$$

fórmula que, junto con [15], proporciona la desigualdad

$$\left|\int_a^b f(x)\,dx\right| \leqslant \int_a^b |f(x)|\,dx. \qquad [16]$$

II. LA DERIVADA

1. La derivada como pendiente.—Mientras que el concepto de integral tiene sus raíces en la antigüedad clásica, la otra idea fundamental del cálculo, la derivada, no se formuló hasta el siglo XVII por Fermat y otros. Fué el descubrimiento, efectuado por Newton y Leibniz, de la relación orgánica entre estas dos ideas, aparentemente tan dispares, lo que inició el inigualado desarrollo de la ciencia matemática.

Fermat estaba interesado por el problema de encontrar los máximos y mínimos de una función $y = f(x)$. En la gráfica, un máximo corresponde a una cúspide, más alta que todos los demás puntos próximos, mientras un mínimo corresponde al fondo de un valle situado por debajo de todos los puntos inmediatos. En la figura 191 el punto B

es un máximo, y el C, un mínimo. Para caracterizar los puntos máximos o mínimos es natural utilizar el concepto de *tangente* a una curva. Supongamos que la gráfica no tiene puntos angulosos u otras singularidades, y que en todo punto posee una dirección definida, dada por una recta tangente. En los puntos de máximo o mínimo, la tangente a la gráfica de la curva $y = f(x)$ debe ser paralela al eje de las x, puesto que, de lo contrario, la curva ascendería o descendería en dichos puntos. Esta observación nos sugiere la idea de considerar con toda generalidad, en cualquier punto, P, de la gráfica $y = f(x)$, la dirección de la tangente a la curva.

Para determinar la dirección de una recta en el plano (x, y), es costumbre dar su *pendiente*, que es la tangente trigonométrica del ángulo de la dirección del semieje positivo de las x con la recta. Si P es un

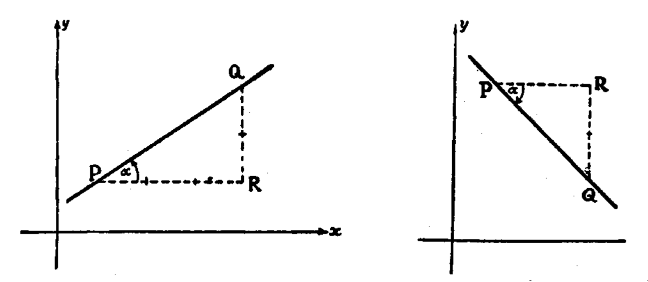


FIG. 267.—Pendientes de rectas.

punto cualquiera de la recta L, nos dirigimos hacia la derecha hasta un punto, R, y después hacia arriba o hacia abajo hasta otro punto, Q, de la recta; por ello, la pendiente de L será $\operatorname{tg} \alpha = RQ/PR$. La longitud PR se toma en sentido positivo, mientras que se considera que RQ es positiva o negativa, según haya que subir o bajar para pasar de R a Q, por lo que la pendiente da la subida o el descenso por unidad de longitud a lo largo de la horizontal cuando nos desplazamos sobre la recta de izquierda a derecha. En la figura 267, la pendiente de la primera recta es $^2/_3$, y la de la segunda -1.

Por pendiente de una *curva* en un punto P, entendemos la pendiente de la tangente a la curva en P. Mientras aceptemos la tangente a una curva como un concepto matemático dado intuitivamente, el único problema que se plantea es el de *hallar un procedimiento para calcular la pendiente*. Por el momento aceptaremos este punto de vista, dejando para más adelante (véase suplemento a este capítulo) un análisis más profundo del problema en cuestión.

2. La derivada como límite.—La pendiente de una curva $y=f(x)$ en el punto $P(x, y)$ no puede calcularse con referencia exclusiva a la curva en el punto P, sino que tenemos que recurrir a un paso al límite, bastante análogo al utilizado para calcular el área bajo una curva, y este paso al límite constituye la base del cálculo diferencial. Consideremos en la curva otro punto, P_1, próximo al P, de coordenadas x_1, y_1. Llamaremos t_1 a la recta que une P con P_1; es una secante de la curva que se aproxima a la tangente a la curva en P cuando P_1 está muy próximo a P. Designaremos por α_1 el ángulo que forma t_1 con el eje de las x. Si hacemos tender ahora x_1 a x, P_1 se moverá a lo largo de la curva hacia P, y la secante t_1 tenderá, como posición lí-

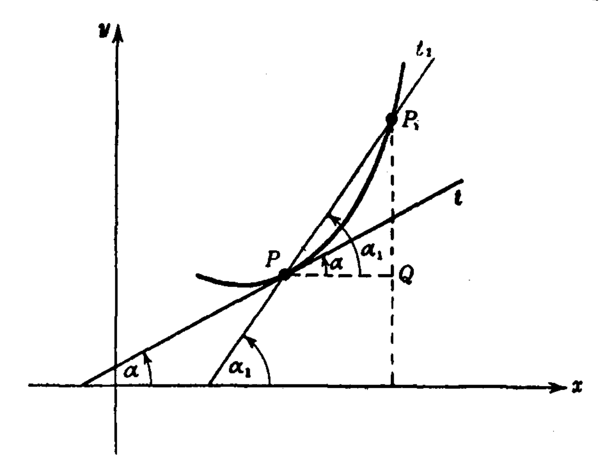


FIG. 268.—La derivada como límite.

mite, a la *tangente* t en el punto P de la curva. Si se designa por α el ángulo que forma t con el eje de las x, al tender x_1 hacia x *, resulta

$$y_1 \rightarrow y, \qquad P_1 \rightarrow P, \qquad t_1 \rightarrow t, \qquad \text{y} \qquad \alpha_1 \rightarrow \alpha.$$

La tangente es el límite de la secante, y la pendiente de la tangente es el límite de la pendiente de la secante.

Aunque no tenemos una expresión explícita de la pendiente de la propia tangente t, sabemos que la pendiente de la secante t_1 está dada por la fórmula

$$\text{pendiente de } t_1 = \frac{y_1 - y}{x_1 - x} = \frac{f(x_1) - f(x)}{x_1 - x},$$

* La notación empleada aquí es ligeramente distinta de la utilizada en el capítulo VI, pues allí considerábamos fijo x_1 en $x \rightarrow x_1$. No debe haber ninguna confusión por razón de esta permutación de símbolos.

o bien, si designamos nuevamente la operación de formar una diferencia mediante el símbolo Δ,

$$\text{pendiente de } t_1 = \frac{\Delta y}{\Delta x} = \frac{\Delta f(x)}{\Delta x}$$

La pendiente de la secante t_1 es un «cociente de diferencias», o sea, la diferencia Δy de los valores de la función dividida por la diferencia Δx de los valores de la variable independiente. Además, pend. de $t =$ $= \text{lím de la pend. de } t_1 = \text{lím} \dfrac{f(x_1) - f(x)}{x_1 - x} = \text{lím} \dfrac{\Delta y}{\Delta x}$, donde los límites se toman suponiendo que $x_1 \to x$; es decir, que $\Delta x = x_1 - x \to 0$. *La pendiente de la tangente* t *a la curva es el límite del cociente de diferencias* $\Delta y/\Delta x$ *cuando* $\Delta x = x_1 - x$ *tiende a cero.*

La función dada $f(x)$ proporciona la *altura* de la curva $y = f(x)$ para el valor x. Podemos considerar ahora la *pendiente* de la curva en un punto variable, P, de coordenadas x, y $[=f(x)]$, como una nueva función de x, que representaremos por $f'(x)$ y llamaremos *derivada* de la función $f(x)$. Se llama *derivación* de $f(x)$ al paso al límite mediante el cual se la ha obtenido. Este proceso es una operación que asocia a una función dada $f(x)$, otra función $f'(x)$, de acuerdo con una regla definida, exactamente igual que la función $f(x)$ viene definida por una regla que atribuye a un valor cualquiera de la variable x el valor $f(x)$:

$f(x)$ = altura de la curva $y = f(x)$ en el punto x;
$f'(x)$ = pendiente de la curva $y = f(x)$ en el punto x.

La derivada $f'(x)$ es, pues, el límite del cociente de la diferencia $f(x_1) - f(x)$ por la diferencia $x_1 - x$:

$$f'(x) = \text{lím} \frac{f(x_1) - f(x)}{x_1 - x} \text{ cuando } x_1 \to x. \qquad [1]$$

Otra notación, a menudo útil, es:

$$f'(x) = Df(x),$$

donde la D es una simple abreviatura de «derivada de». Todavía otra notación distinta es la empleada por Leibniz para designar la derivada de $y = f(x)$:

$$\frac{dy}{dx} \quad \text{o} \quad \frac{df(x)}{dx},$$

que estudiaremos más adelante, y que indica el carácter de la derivada como límite del cociente de diferencias o incrementos $\Delta y/\Delta x$ o $\Delta f(x)/\Delta x$.

Si recorremos la curva $y = f(x)$ en la dirección que corresponde a valores crecientes de x, la *derivada positiva*, $f'(x) > 0$, en un punto significa que *la curva se eleva* (el valor de y aumenta), y al contrario, una *derivada negativa*, $f'(x) < 0$, quiere decir que *la curva desciende*, mientras que $f'(x) = 0$ significa que para ese valor de x la curva tiene una dirección horizontal. En un máximo o mínimo, la pendiente debe ser igual a cero (Fig. 269).

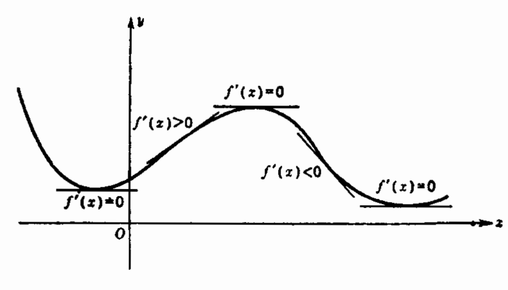


Fig. 269.—Signo de la derivada.

De ahí que resolviendo la ecuación $f'(x) = 0$ respecto de x, se puede determinar la posición de los máximos y mínimos, como lo hizo Fermat por primera vez.

3. Ejemplos.—Podría parecer que las observaciones que nos condujeron a la definición [1] carecen de valor práctico. Se ha reemplazado un problema por otro; en lugar de preguntar cuál es la pendiente de la tangente a una curva $y = f(x)$ en un punto, se pregunta por el valor de un límite [1], lo que, a primera vista, parece igualmente difícil. Pero en cuanto abandonamos el dominio de las generalidades y consideramos funciones particulares $f(x)$ obtenemos resultados tangibles.

La función más sencilla es $f(x) = c$, donde c es una constante. La gráfica de la función $y = f(x) = c$ es una recta horizontal que coincide con todas sus tangentes, y es evidente que $f'(x) = 0$ para todos los valores de x. Esto se deduce también de la definición [1], pues

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x} = \frac{c - c}{x_1 - x} = \frac{0}{x_1 - x} = 0,$$

por lo que resulta trivial que

$$\lim \frac{f(x_1) - f(x)}{x_1 - x} = 0 \quad \text{cuando} \quad x_1 \to x.$$

Como segundo caso consideraremos la función $y = f(x) = x$, cuya gráfica es una recta que pasa por el origen y es bisectriz del primer cuadrante. Geométricamente es obvio que

$$f'(x) = 1$$

para todos los valores de x, y la definición analítica [1] proporciona también el mismo resultado:

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1 - x}{x_1 - x} = 1,$$

de modo que

$$\text{lím } \frac{f(x_1) - f(x)}{x_1 - x} = 1 \quad \text{para} \quad x_1 \to x.$$

El ejemplo más sencillo y ya no trivial nos lo proporciona la derivación de la función

$$y = f(x) = x^2,$$

lo que equivale a encontrar la pendiente de una parábola. Este es el caso más simple que nos enseña cómo se ha de efectuar el paso al límite cuando el resultado no es evidente de antemano. Tenemos:

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1^2 - x^2}{x_1 - x}$$

Si intentáramos efectuar directamente el paso al límite en numerador y denominador, obtendríamos la expresión carente de significado 0/0. Pero podemos evitar este escollo eliminando *antes de pasar al límite* el factor perturbador $x_1 - x$. (Al calcular el límite del cociente de diferencias, consideramos exclusivamente valores $x_1 \neq x$, por lo que está permitido efectuar esa simplificación; véase pág. 318). Así obtenemos la expresión:

$$\frac{x_1^2 - x^2}{x_1 - x} = \frac{(x_1 - x)(x_1 + x)}{x_1 - x} = x_1 + x.$$

Después de la simplificación, no existe ninguna otra dificultad con el límite, cuando $x_1 \to x$. Este se obtiene «por sustitución», pues en la nueva forma, $x_1 + x$, el cociente de diferencias es función continua, y el límite de una función continua para $x_1 \to x$ es simplemente el valor de la función para $x_1 = x$; o sea, en nuestro caso, $x + x = 2x$, por lo que

$$f'(x) = 2x \quad \text{para} \quad f(x) = x^2.$$

De manera análoga, podemos demostrar que para $f(x) = x^3$, $f'(x) = 3x^2$, pues el cociente de diferencias

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x} = \frac{x_1^3 - x^3}{x_1 - x},$$

puede simplificarse teniendo en cuenta la fórmula $x_1^3 - x^3 = (x_1 - x)(x_1^2 + x_1x + x^2)$, eliminándose el denominador $\Delta x = x_1 - x$, con lo que se obtiene la expresión continua

$$\frac{\Delta y}{\Delta x} = x_1^2 + x_1x + x^2.$$

Si ahora se hace tender x_1 a x, esta expresión tiende a $x^2+x^2+x^2$, de donde resulta $f'(x) = 3x^2$.

En general, para $f(x) = x^n$, siendo n un número entero y positivo cualquiera, se obtiene la derivada

$$f'(x) = nx^{n-1}.$$

Ejercicio: Demuéstrese el resultado anterior. [Utilícese la fórmula algebraica

$$x_1^n - x^n = (x_1 - x)(x_1^{n-1} + x_1^{n-2}x + x_1^{n-3}x^2 + \cdots + x_1x^{n-2} + x^{n-1}).]$$

Como un ejemplo más de los sencillos artificios que permiten determinar explícitamente la derivada, consideremos la función

$$y = f(x) = \frac{1}{x}$$

Tenemos:

$$\frac{\Delta y}{\Delta x} = \frac{y_1 - y}{x_1 - x} = \left(\frac{1}{x_1} - \frac{1}{x}\right)\cdot\frac{1}{x_1 - x} = \frac{x - x_1}{x_1x}\cdot\frac{1}{x_1 - x}$$

Después de simplificar, resulta $\Delta y/\Delta x = -1/x_1x$, que es continua para $x_1 = x$, por lo que en el límite

$$f'(x) = -\frac{1}{x^2}$$

Naturalmente, ni la derivada ni la propia función están definidas para $x = 0$.

Ejercicios: Demuéstrese en forma análoga que para $f(x) = \dfrac{1}{x^2}$, $f'(x) = -\dfrac{2}{x^3}$; para $f(x) = \dfrac{1}{x^n}$, $f'(x) = -\dfrac{n}{x^{n+1}}$; para $f(x) = (1 + x)^n$, $f'(x) = n(1 + x)^{n-1}$.

Vamos a efectuar la derivación de

$$y = f(x) = \sqrt{x}$$

El cociente de diferencias en este caso es:

$$\frac{y_1 - y}{x_1 - x} = \frac{\sqrt{x_1} - \sqrt{x}}{x_1 - x}$$

Mediante la fórmula $x_1 - x = (\sqrt{x_1} - \sqrt{x})\ (\sqrt{x_1} + \sqrt{x})$, podemos eliminar un factor y obtener la expresión continua:

$$\frac{y_1 - y}{x_1 - x} = \frac{1}{\sqrt{x_1} + \sqrt{x}}$$

Pasando al límite se tiene:

$$f'(x) = \frac{1}{2\sqrt{x}}$$

Ejercicios: Demuéstrese que para

$f(x) = \frac{1}{\sqrt{x}}$, $f'(x) = -\frac{1}{2(\sqrt{x})^3}$; para $f(x) = \sqrt[3]{x}$, $f'(x) = \frac{1}{3\sqrt[3]{x^2}}$; para

$f(x) = \sqrt{1 - x^2}$, $f'(x) = \frac{-x}{\sqrt{1 - x^2}}$; para $f(x) = \sqrt[n]{x}$, $f'(x) = \frac{1}{n\sqrt[n]{x^{n-1}}}$

4. Derivadas de las funciones trigonométricas.—Trataremos ahora una cuestión muy importante: *la derivación de las funciones trigonométricas*, para lo cual utilizaremos exclusivamente la medida de los ángulos en radianes.

Para derivar la función $y = f(x) = \operatorname{sen} x$, ponemos $x_1 - x = h$, con lo que resulta $x_1 = x + h$ y $f(x_1) = \operatorname{sen} x_1 = \operatorname{sen}(x + h)$. Mediante la fórmula trigonométrica para sen $(A + B)$, resulta:

$$f(x_1) = \operatorname{sen}(x + h) = \operatorname{sen} x \cos h + \cos x \operatorname{sen} h.$$

De donde,

$$\frac{f(x_1) - f(x)}{x_1 - x} = \frac{\operatorname{sen}(x + h) - \operatorname{sen} x}{h} = \cos x\left(\frac{\operatorname{sen} h}{h}\right) + \operatorname{sen} x\left(\frac{\cos h - 1}{h}\right). \quad [2]$$

Si hacemos tender ahora x_1 a x, h tiende a 0, sen h a 0 y cos h a 1. Además, de acuerdo con los resultados de la página 319, se tiene:

$$\lim \frac{\operatorname{sen} h}{h} = 1 \quad \text{y} \quad \lim \frac{\cos h - 1}{h} = 0.$$

Por tanto, el primer miembro de [2] tiende a cos x, de donde resulta:

La derivada de la función f(x) = sen x *es* f'(x) = cos x, o más brevemente:

$$D \text{ sen } x = \cos x.$$

Ejercicio: Demuéstrese que $D \cos x = -\text{sen } x$.

Para derivar $f(x) = \text{tg } x$, escribiremos $\text{tg } x = \text{sen } x/\cos x$, con lo que resulta:

$$\frac{f(x+h)-f(x)}{h} = \left(\frac{\text{sen }(x+h)}{\cos(x+h)} - \frac{\text{sen } x}{\cos x}\right)\frac{1}{h}$$

$$= \frac{\text{sen }(x+h)\cos x - \cos(x+h)\text{ sen } x}{h}\,\frac{1}{\cos(x+h)\cos x}$$

$$= \frac{\text{sen } h}{h}\,\frac{1}{\cos(x+h)\cos x}$$

[La última igualdad se deduce de la fórmula $\text{sen }(A-B) = \text{sen } A \cos B - \cos A \text{ sen } B$, para $A = x + h$ y $B = h$.] Si hacemos tender ahora h a cero, $\text{sen } h/h$ tiende a 1, $\cos(x+h)$ tiende a $\cos x$, y deducimos:

La derivada de la función f(x) = tg x *es* f'(x) = 1/cos²x, o sea,

$$D \text{ tg } x = \frac{1}{\cos^2 x}$$

Ejercicio: Demuéstrese que $D \cot x = -1/\text{sen}^2 x$.

***5. Derivación y continuidad.**—Vamos a demostrar que *toda función derivable es continua.* En efecto, si existe el límite de $\Delta y/\Delta x$ al tender Δx a cero, es fácil ver que la diferencia Δy de la función $f(x)$ llega a ser tan pequeña como se desee al tender a cero la diferencia Δx. De ahí que siempre que pueda derivarse una función, su continuidad está automáticamente asegurada; por consiguiente prescindiremos de mencionar explícitamente o de demostrar la continuidad de las funciones derivables que aparezcan en este capítulo, a no ser que exista una razón particular para ello.

6. Derivada y velocidad. Segunda derivada y aceleración.—En lo que precede se ha estudiado la derivada desde el punto de vista geométrico de la gráfica de la función. Pero el significado del concepto de derivada no se limita al problema de hallar la pendiente de la tangente a una curva. En las ciencias naturales importa mucho más determinar la variación relativa de cierta magnitud $f(t)$ que varía con el tiempo t. Fué en esta dirección por la cual llegó Newton al cálculo diferencial. Deseaba, en particular, analizar el fenómeno de la veloci-

dad, considerando el tiempo y la posición de una partícula en movimiento como variables, o, como decía Newton, «cantidades fluyentes».

Si una partícula se mueve siguiendo una trayectoria rectilínea, su movimiento está completamente descrito dando la posición x para un tiempo cualquiera t como una función $x = f(t)$. Se define un «movimiento uniforme» de velocidad constante b a lo largo del eje x mediante una función lineal $x = a + bt$, donde a es la abscisa de la partícula en el tiempo $t = 0$.

En el plano, el movimiento de una partícula queda descrito por dos funciones:

$$x = f(t), \qquad y = g(t),$$

que determinan las coordenadas en función del tiempo. En particular, un movimiento uniforme corresponde a un par de funciones lineales

$$x = a + bt, \qquad y = c + dt,$$

donde b y d son las dos «componentes» de la velocidad constante, y a y c, las coordenadas de la partícula en el instante $t = 0$; la trayectoria de la partícula es una recta, de ecuación

$$(x - a)d - (y - c)\,b = 0,$$

que se obtiene por eliminación de t entre las dos ecuaciones anteriores.

Si una partícula se mueve en el plano vertical x, y, exclusivamente sometida a la acción de la gravedad, según se demuestra en física elemental, su movimiento está determinado por las dos ecuaciones siguientes:

$$x = a + bt, \qquad y = c + dt - {}^1/{}_2gt^2,$$

donde a, b, c, d, son constantes que dependen del estado inicial de la partícula, y g la aceleración debida a la gravedad (aproximadamente igual a 9,80 m/seg^2 si la distancia se mide en metros y el tiempo en segundos). La trayectoria de la partícula, obtenida eliminando t entre ambas ecuaciones, es una parábola

$$y = c + \frac{d}{b}(x - a) - {}^1/{}_2g\,\frac{(x - a)^2}{b^2},$$

si $b \neq 0$; en otro caso es un segmento del eje vertical.

Si una partícula se mueve a lo largo de una curva del plano (como un tren sobre la vía), su movimiento puede describirse dando la longitud del arco s, medido a partir de un punto inicial, P_0, hasta la posición P de la partícula en el instante t, como función de t: $s = f(t)$; p. ej., en el círculo unidad $x^2 + y^2 = 1$, la función $s = ct$ describe una rotación uniforme de velocidad c, a lo largo de la circunferencia.

Ejercicios: *Dibújense las trayectorias de los movimientos planos descritos por:
1. $x = \operatorname{sen} t,\ y = \cos t.$
2. $x = \operatorname{sen} 2t,\ y = \operatorname{sen} 3t.$
3. $x = \operatorname{sen} 2t,\ y = 2 \operatorname{sen} 3t.$
4. En el movimiento parabólico descrito anteriormente, supongamos que la partícula se encuentra en el origen para $t = 0$ y $b > 0$, $d > 0$. Hállense las cordenadas del punto más alto de la trayectoria. Determínense el tiempo, t, y el valor de x para la segunda intersección de la trayectoria con el eje de las x.

El propósito principal de Newton fué el de determinar la velocidad en un movimiento no uniforme. Para mayor sencillez, consideremos el movimiento de una partícula sobre una recta, definido por la función $x = f(t)$. Si el movimiento fuera uniforme, con velocidad constante, podría hallarse la velocidad tomando dos valores t y t_1 del tiempo y los correspondientes valores $x = f(t)$ y $x_1 = f(t_1)$, y formando el cociente:

$$v = \text{velocidad} = \frac{\text{distancia}}{\text{tiempo}} = \frac{x_1 - x}{t_1 - t} = \frac{f(t_1) - f(t)}{t_1 - t}$$

Así, p. ej., si t se mide en horas y x en kilómetros, para $t_1 - t = 1$, $x_1 - x$ será el número de kilómetros recorridos en una hora, y v será la velocidad en kilómetros por hora. Decir que la velocidad del móvil es constante equivale simplemente a que el cociente de diferencias

$$\frac{f(t_1) - f(t)}{t_1 - t} \qquad [3]$$

es el mismo para todos los valores de t y t_1. Pero cuando el movimiento no es uniforme, como ocurre en el caso de un grave, cuya velocidad aumenta con el tiempo, el cociente [3] no da la velocidad *en el instante t*, sino la *velocidad media* durante el intervalo de tiempo de t a t_1. Para obtener la velocidad en el instante preciso t, debemos hallar el límite de la velocidad media cuando t_1 tiende a t. Así, definiremos con Newton:

$$\text{velocidad en el instante } t = \lim \frac{f(t_1) - f(t)}{t_1 - t} = f'(t). \qquad [4]$$

En otras palabras, la velocidad es la derivada de la distancia respecto al tiempo o «la variación instantánea» de la distancia respecto al tiempo (lo que es distinto de la variación *media* definida por [3]).

Se llama *aceleración* a la *variación unitaria de la velocidad.* Es simplemente la derivada de la derivada, que se representa generalmente por $f''(t)$ y se llama *segunda derivada* de $f(t)$.

Galileo observó que para un cuerpo que cae libremente, la distancia vertical que recorre durante el tiempo t está dada por la fórmula

$$x = f(t) = {}^1/_2 gt^2, \qquad [5]$$

donde g es la constante de gravitación. Derivando [5], se deduce que la velocidad, v, del cuerpo en el instante, t, está dada por

$$v = f'(t) = gt, \qquad [6]$$

y la aceleración, α, por

$$\alpha = f''(t) = g,$$

que es constante.

Supongamos que se desea calcular la velocidad del cuerpo, 2 segundos después de iniciar su caída. La velocidad *media* durante el intervalo de tiempo de $t = 2$ a $t = 2,1$ es:

$$\frac{{}^1/_2 g(2,1)^2 - {}^1/_2 g(2)^2}{2,1 - 2} = 20,09 \text{ (metros por segundo).}$$

Sustituyendo $t = 2$ en [6], encontramos la velocidad *instantánea* al final de los dos segundos, que es $v = 19,6$.

Ejercicio: ¿Cuál es la velocidad media del cuerpo durante el intervalo de tiempo de $t = 2$ a $t = 2,01$, y de $t = 2$ a $t = 2,001$?

En el caso de un movimiento plano, las dos derivadas $f'(t)$ y $g'(t)$ de las funciones $x = f(t)$, $y = g(t)$ definen las componentes de la velocidad. Para el movimiento a lo largo de una curva dada, la velocidad estará definida por la derivada de la función $s = f(t)$, donde s es la longitud del arco.

7. Significado geométrico de la segunda derivada.—La segunda derivada es también importante en geometría y análisis, puesto que $f''(x)$ expresa la variación unitaria de la pendiente $f'(x)$ de la curva $y = f(x)$, por lo que sirve de indicación de cómo se modifica la forma de la curva. Si $f''(x)$ es positiva en un intervalo, entonces la variación relativa de $f'(x)$ es positiva, lo que significa que la función crece al aumentar x. En consecuencia, $f''(x) > 0$ significa que la pendiente $f'(x)$ aumenta al crecer x, por lo que la curva se acerca más a la vertical cuando la pendiente es positiva, y se separa de ella cuando es negativa. Decimos entonces que la curva tiene su *concavidad dirigida hacia arriba* (Fig. 270).

Análogamente, si $f''(x) < 0$, la curva $y = f(x)$ tiene su *concavidad dirigida hacia abajo* (Fig. 271).

La parábola $y = f(x) = x^2$ tiene su concavidad dirigida constantemente hacia arriba, puesto que $f''(x) = 2$ es siempre positiva. La

curva $y = f(x) = x^3$ tiene su concavidad dirigida hacia arriba para $x > 0$, y hacia abajo para $x < 0$ (Fig. 153), pues $f''(x) = 6x$, como el lector puede comprobar fácilmente. Para $x = 0$, se tiene $f'(x) = 3x^2 = 0$

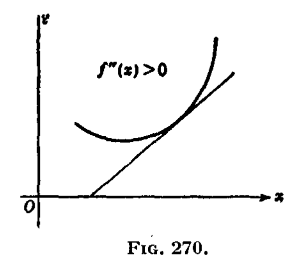


FIG. 270.

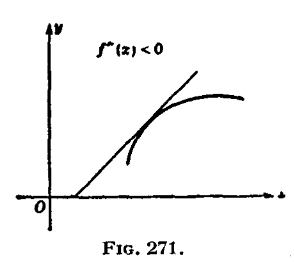


FIG. 271.

(sin máximo ni mínimo); también $f''(x) = 0$ para $x = 0$. Este punto se denomina *punto de inflexión*. En tales puntos, la tangente (en este caso el eje x) atraviesa la curva.

Si s representa la longitud del arco y α el ángulo de pendiente, $\alpha = h(s)$ será una función de s, que variará al recorrer la curva. La derivada $h'(s)$ se llama *curvatura* de la curva en el punto donde el arco tiene longitud s. Sin detenernos a dar la correspondiente demostración, recordaremos que la curvatura $\varkappa$ puede expresarse mediante la primera y la segunda derivadas de la función $y = f(x)$ que define la curva:

$$\varkappa = f''(x) / [1 + (f'(x))^2]^{3/2}.$$

8. Máximos y mínimos.—Podemos determinar los máximos y mínimos de una función dada $f(x)$, calculando la primera derivada $f'(x)$; hallando los valores que la anulan, e investigando, finalmente, cuáles de ellos corresponden a máximos o a mínimos. Puede decidirse esta última cuestión formando la segunda derivada $f''(x)$, cuyo signo nos dice si la curva es cóncava o convexa, y cuya anulación indica en general la existencia de un punto de inflexión en el cual no existe extremo. Mediante los signos de $f'(x)$ y $f''(x)$ podemos determinar no sólo los extremos, sino la forma de la gráfica de la función. Este método nos da los valores de x para los cuales hay máximo o mínimo; los valores correspondientes de $y = f(x)$ se tendrán sustituyendo estos valores de x en $f(x)$.

Como ejemplo, consideraremos el polinomio

$$f(x) = 2x^3 - 9x^2 + 12x + 1,$$

para el cual se tiene:

$$f'(x) = 6x^2 - 18x + 12, \qquad f''(x) = 12x - 18.$$

Las raíces de la ecuación de segundo grado $f'(x) = 0$ son $x_1 = 1$ y $x_2 = 2$; y tenemos: $f''(x_1) = -6 < 0$, $f''(x_2) = 6 > 0$. Por tanto, la función $f(x)$ tiene un máximo $f(x_1) = 6$ y un mínimo $f(x_2) = 5$.

Ejercicios:

1. **Trácese la gráfica de la función anterior.**
2. **Discútase y trácese la gráfica de $f(x) = (x^2 - 1)(x^2 - 4)$.**
3. **Determínese el mínimo de $x + 1/x$; de $x + a^2/x$; de $px + q/x$, siendo p y q positivos. ¿Tienen máximos estas funciones?**
4. **Determínense los máximos y mínimos de sen x y sen (x^2).**

III. TÉCNICA DE LA DERIVACIÓN

Hasta ahora nuestros esfuerzos se han dirigido a derivar diversas funciones transformando adecuadamente el cociente de incrementos antes de realizar el paso al límite. Se dió un paso decisivo cuando, debido a las investigaciones de Newton y Leibniz y sus sucesores, se reemplazaron esos procedimientos particulares por métodos generales más potentes. Mediante ellos, se puede derivar casi automáticamente cualquier función de las que normalmente aparecen en matemáticas, una vez dominadas unas pocas reglas muy sencillas y aprendido a reconocer su aplicabilidad. Así, la derivación ha adquirido el carácter de un «algoritmo» de cálculo, y es este aspecto de la teoría el que se expresa mediante la palabra «cálculo».

No podemos dar muchos detalles de esta técnica, y sólo mencionaremos unas cuantas reglas.

a) *Derivación de una suma.* Si a y b son constantes y se define la función $k(x)$ por

$$k(x) = af(x) + bg(x),$$

se tiene, como el lector puede verificar fácilmente,

$$k'(x) = af'(x) + bg'(x).$$

Una regla análoga es válida para cualquier número de términos.

b) *Derivación de un producto.* Para un producto

$$p(x) = f(x)g(x),$$

la derivada es:

$$p'(x) = f(x)g'(x) + g(x)f'(x).$$

Esto se demuestra fácilmente mediante el siguiente artificio, que consiste en sumar y restar un mismo término:

$$\begin{aligned} p(x + h) - p(x) &= f(x + h)g(x + h) - f(x)g(x) = \\ &= f(x + h)g(x + h) - f(x + h)g(x) + f(x + h)g(x) - f(x)g(x); \end{aligned}$$

de aquí se obtiene:

$$\frac{p(x+h)-p(x)}{h} = f(x+h)\frac{g(x+h)-g(x)}{h} + g(x)\frac{f(x+h)-f(x)}{h}$$

Hagamos tender ahora h a cero; puesto que $f(x+h)$ tiende a $f(x)$, la regla queda inmediatamente establecida.

Ejercicio: Demuéstrese que la función $p(x) = x^n$ tiene la derivada $p'(x) = nx^{n-1}$. (Escríbase $x^n = x \cdot x^{n-1}$ y utilícese el principio de inducción.)

Utilizando las reglas *a*) y *b*) podemos derivar cualquier polinomio:

$$f(x) = a_0 + a_1x + \cdots + a_nx^n;$$

su derivada es:

$$f'(x) = a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1}.$$

Como aplicación, podemos demostrar el *teorema del binomio* (véase pág. 24). Este teorema se refiere al desarrollo de $(1+x)^n$ en forma de polinomio:

$$f(x) = (1+x)^n = 1 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n, \qquad [1]$$

y dice que el coeficiente a_k está dado por la fórmula

$$a_k = \frac{n(n-1)\cdots(n-k+1)}{k!}, \qquad [2]$$

siendo, naturalmente, $a_n = 1$.

Ya hemos visto (pág. 430, *Ejercicio*) que la derivada del primer miembro de [1] es $n(1+x)^{n-1}$; pero, según el párrafo anterior, se tiene:

$$n(1+x)^{n-1} = a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1}. \qquad [3]$$

Hagamos $x = 0$ en esta fórmula, de donde resulta $n = a_1$, que es precisamente el valor de [2] para $k = 1$. Derivando de nuevo [3], se tiene:

$$n(n-1)(1+x)^{n-2} = 2a_2 + 3 \cdot 2a_3x + \cdots + n(n-1)a_nx^{n-2},$$

y haciendo $x = 0$, resulta $n(n-1) = 2a_2$, de acuerdo con la fórmula [2] para $k = 2$.

Ejercicio: Demuéstrese [2] para $k = 3$, 4, y para un k cualquiera mediante el método de inducción matemática.

c) *Derivada de un cociente.* Si es

$$q(x) = \frac{f(x)}{g(x)},$$

se tiene:

$$q'(x) = \frac{g(x)f'(x) - f(x)g'(x)}{[g(x)]^2}$$

La demostración queda como ejercicio a cargo del lector.

Ejercicio: Mediante esta regla, obténganse las fórmulas de la página 432 para las derivadas de tg x y cot x, a partir de las de sen x y cos x. Demuéstrese que las derivadas de sec $x = 1/\cos x$ y cosec $x = 1/\text{sen } x$ son. resnectivamente, sen $x/\cos^2 x$ y, $-\cos x/\text{sen}^2 x$.

Podemos ahora derivar cualquier función que esté definida como cociente de dos polinomios; p. ej.,

$$f(x) = \frac{1 - x}{1 + x}$$

tiene la derivada

$$f'(x) = \frac{-(1 + x) - (1 - x)}{(1 + x)^2} = -\frac{2}{(1 + x)^2}$$

Ejercicio: Derívese la función: $f(x) = 1/x^m = x^{-m}$, donde m es un entero positivo. El resultado es:

$$f'(x) = -mx^{-m-1},$$

d) *Derivación de funciones inversas.* Si

$$y = f(x) \qquad \text{y} \qquad x = g(y)$$

son funciones inversas (p. ej., $y = x^2$ y $x = \sqrt{y}$), sus derivadas son recíprocas:

$$g'(y) = \frac{1}{f'(x)} \qquad \text{o} \qquad Dg(y) \cdot Df(x) = 1.$$

Se demuestra fácilmente este hecho recurriendo a los cocientes de diferencias: $\Delta y/\Delta x$ y $\Delta x/\Delta y$, respectivamente. Puede deducirse también de la interpretación geométrica de la función inversa dada en la página 292, si referimos la pendiente de la tangente al eje de las y en lugar de al de las x.

Como ejemplo, derivaremos la función

$$y = f(x) = \sqrt[m]{x} = x^{\frac{1}{m}},$$

que es la inversa de $x = y^m$ (véase también el estudio más directo de esta cuestión para $m = 1/2$, hecho en la pág. 431). Puesto que la derivada de la última función es my^{m-1}, se tiene:

$$f'(x) = \frac{1}{my^{m-1}} = \frac{1}{m}\,\frac{y}{y^m} = \frac{1}{m}\,yy^{-m},$$

de donde, sustituyendo $y = x^{\frac{1}{m}}$ y $y^{-m} = x^{-1}$, $f'(x) = \frac{1}{m}\,x^{\frac{1}{m}-1}$, o bien,

$$D(x^{1/m}) = \frac{1}{m}\,x^{\frac{1}{m}-1}$$

Como otro ejemplo más, derivaremos las *funciones trigonométricas inversas* (véase pág. 293):

$$y = \text{arc tg } x, \qquad \text{que equivale a} \qquad x = \text{tg } y.$$

En este caso, la variable y, que se da expresada en radianes, está limitada al intervalo $-1/2\pi < y < 1/2\pi$ para asegurar la univocidad de la función inversa.

Según hemos visto (pág 432), $D \text{ tg } y = 1/\cos^2 y$, y dado que $1/\cos^2 y = (\text{sen}^2 y + \cos^2 y)/\cos^2 y = 1 + \text{tg}^2 y = 1 + x^2$, se tiene:

$$D \text{ arc tg } x = \frac{1}{1 + x^2}$$

De la misma manera puede deducir el lector las siguientes fórmulas:

$$D \text{ arc cot } x = -\frac{1}{1 + x^2};$$

$$D \text{ arc sen } x = \frac{1}{\sqrt{1 - x^2}};$$

$$D \text{ arc cos } x = -\frac{1}{\sqrt{1 - x^2}}$$

Llegamos por fin a una regla general muy importante:

e) *Derivación de las funciones compuestas.* Tales funciones están formadas por otras dos (o más) que son más sencillas (véase pág. 293); p. ej., $z = \text{sen}(\sqrt{x})$ se compone de $z = \text{sen } y$ e $y = \sqrt{x}$; la función $z = \sqrt{x} + \sqrt{x^5}$ se compone de $z = y + y^5$ e $y = \sqrt{x}$; $z = \text{sen } (x^2)$ se compone de $z = \text{sen } y$ e $y = x^2$; $z = \text{sen } 1/x$ se compone de $z = \text{sen } y$ e $y = 1/x$.

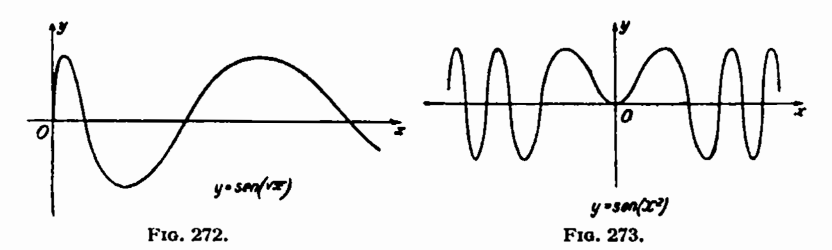


FIG. 272. FIG. 273.

Si se dan dos funciones:

$$z = g(y) \quad \text{e} \quad y = f(x),$$

y se sustituye la última en la primera, obtenemos una función compuesta:

$$z = k(x) = g[f(x)] .$$

Pues bien: vamos a ver que

$$k'(x) = g'(y)f'(x). \qquad [4]$$

En efecto, si escribimos:

$$\frac{k(x_1) - k(x)}{x_1 - x} = \frac{z_1 - z}{y_1 - y} \cdot \frac{y_1 - y}{x_1 - x},$$

siendo $y_1 = f(x_1)$ y $z_1 = g(y_1) = k(x_1)$, y hacemos tender x_1 a x, el primer miembro tiende a $k'(x)$, mientras que los dos factores del segundo miembro tienden a $g'(x)$ y $f'(x)$, respectivamente, de acuerdo con [4].

En esta demostración es necesario suponer que $y_1 - y \neq 0$, ya que dividimos por $\Delta y = y_1 - y$, y no podemos utilizar valores de x_1 para los cuales $y_1 - y = 0$. Pero la fórmula [4] conserva su validez aun cuando $\Delta y = 0$ en un entorno de x; y es entonces constante, $f'(x) = 0$, $k(x) = g(y)$ es constante respecto a x (puesto que y no varía con x), y, en consecuencia, $k'(x) = 0$, como debía ocurrir en este caso según [4]. El lector verificará los ejemplos siguientes:

$$k(x) = \operatorname{sen} \sqrt{x}, \qquad k'(x) = (\cos \sqrt{x}) \frac{1}{2\sqrt{x}};$$

$$k(x) = \sqrt{x} + \sqrt{x^5}, \qquad k'(x) = (1 + 5x^2) \frac{1}{2\sqrt{x}};$$

$$k(x) = \operatorname{sen}(x^2), \qquad k'(x) = \cos(x^2) \cdot 2x;$$

$$k(x) = \operatorname{sen} \frac{1}{x}, \qquad k'(x) = -\cos\left(\frac{1}{x}\right) \frac{1}{x^2};$$

$$k(x) = \sqrt{1 - x^2}, \qquad k'(x) = \frac{-1}{2\sqrt{1 - x^2}} \cdot 2x = \frac{-x}{\sqrt{1 - x^2}}$$

Ejercicio: Combinando los resultados de las páginas 430 y 440, demuéstrese que la función

$$f(x) = \sqrt[m]{x^s} = x^{\frac{s}{m}}$$

tiene por derivada:

$$f'(x) = \frac{s}{m} x^{\frac{s}{m}-1}$$

Debe observarse que todas las fórmulas referentes a potencias de x pueden expresarse en una sola:

Si r *es un número racional cualquiera, positivo o negativo, la función*

$$f(x) = x^r$$

tiene como derivada:

$$f'(x) = rx^{r-1} .$$

Ejercicios:

1. Efectúense las derivaciones de los ejercicios de la página 431, utilizando las reglas que acabamos de establecer.
2. Derívense las siguientes funciones:

$$x \operatorname{sen} x, \frac{1}{1+x^2} \operatorname{sen} nx, (x^3 - 3x^2 - x + 1)^3, 1 + \operatorname{sen}^2 x, x^2 \operatorname{sen} \frac{1}{x^2},$$

$$\operatorname{arc\,sen} (\cos nx), \operatorname{tg} \frac{1+x}{1-x}, \operatorname{arc\,tg} \frac{1+x}{1-x}, \sqrt[4]{1-x^2}, \frac{1}{1+x^2}$$

3. Determínense las segundas derivadas de algunas de las funciones precedentes, y además, de $\operatorname{arc\,tg} x$, $\operatorname{sen}^2 x$, $\operatorname{tg} x$ y $\frac{1-x}{1+x}$.
4. Derívese $c_1 (x - x_1)^2 + y_1^2 + c_2(x - x_2)^2 + y_2^2$, y demuéstrense las propiedades de mínimo del rayo luminoso en la reflexión y refracción, que se expusieron en el capítulo VII (págs. 341 y 392). Tanto la reflexión como la refracción tienen lugar en el eje de las x, y las coordenadas de los puntos extremos de la trayectoria son, respectivamente, x_1, y_1, y x_2, y_2. *(Observación*: La derivada de la función se anula en un solo punto; por tanto, dado que puede existir mínimo, pero no máximo, no es necesario estudiar la segunda derivada.)

Otros problemas de máximos y mínimos:

5. Determínense los extremos de las siguientes funciones, dibújense sus respectivas gráficas y establézcase en qué intervalos la función es creciente, decreciente, cóncava o convexa:

$$x^3 - 6x + 2, \ x/(1 + x^2), \ x^2/(1 + x^4), \ \cos^2 x.$$

6. Estúdiense los máximos y mínimos de la función $x^3 + 3ax + 1$ en su dependencia respecto a a.
7. ¿Qué punto de la hipérbola $2y^2 - x^2 = 2$ es el más próximo al punto $x = 0$, $y = 3$?
8. Entre todos los rectángulos de área dada, determínese aquel cuya diagonal sea mínima.

9. Inscríbase el rectángulo de área máxima en la elipse $x^2/a^2 + y^2/b^2 = 1$.

10. Entre todos los cilindros circulares de volumen dado, determínese el de área mínima.

IV. LA NOTACIÓN DE LEIBNIZ Y «LOS INFINITÉSIMOS»

Newton y Leibniz sabían cómo obtener integrales y derivadas mediante pasos al límite, pero los fundamentos del cálculo se hallaban oscurecidos por la incapacidad para reconocer al concepto de límite el derecho exclusivo como fuente de los nuevos métodos. Por muy sencillo que esto nos parezca ahora, después de haber aclarado completamente el concepto de límite, ni Newton ni Leibniz se atrevieron a adoptar una actitud decidida. Su ejemplo influyó durante más de un siglo sobre el desarrollo de la ciencia matemática, período durante el cual se soslayaba el tema hablando de «cantidades infinitamente pequeñas», «diferenciales», «razón última», etc. La repugnancia con que se abandonaron por fin esos conceptos estaba profundamente arraigada en la actitud filosófica de la época y en la propia naturaleza de la inteligencia humana. Cabe hacerse este razonamiento: Por supuesto, tanto la integral como la derivada se calculan mediante límites; pero, después de todo, ¿qué son estos entes en sí mismos, fuera del particular paso al límite mediante el cual se los describe? Parece evidente que conceptos intuitivos tales como la superficie o la pendiente de una curva tienen un significado absoluto en sí, sin necesidad de la idea auxiliar de los polígonos inscritos o rectas secantes y de sus límites. Psicológicamente resulta muy natural que se intente buscar una definición adecuada de la superficie o de la pendiente, como «cosas en sí». Pero la renuncia a este deseo y la aceptación del paso al límite como única definición aceptable, desde el punto de vista científico, está de acuerdo con la madurez científica que ha abierto tantas veces el camino al progreso. En el siglo XVII no existía una tradición intelectual que permitiese tal radicalismo filosófico.

La tentativa de Leibniz de «explicar» la derivada empezaba correctamente con el cociente de diferencias de una función $y = f(x)$:

$$\frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x)}{x_1 - x}$$

Para designar el límite, esto es, la derivada que nosotros representamos por $f'(x)$ (siguiendo la notación introducida por Lagrange), Leibniz escribía:

$$\frac{dy}{dx},$$

reemplazando el símbolo Δ de diferencias por el «símbolo diferencial» *d*. No se plantea ninguna dificultad ni existe ningún misterio si entendemos bien que este símbolo indica solamente que debe efectuarse el paso al límite: $\Delta x \to 0$ y, en consecuencia, $\Delta y \to 0$. *Antes* de pasar al límite se elimina el denominador Δx del cociente $\frac{\Delta y}{\Delta x}$ o se transforma de tal manera que pueda efectuarse sin dificultad el proceso de límite. Éste es siempre el punto crucial en la derivación. Si hubiéramos intentado hacerlo sin esa simplificación previa, habríamos obtenido la relación carente de sentido $\Delta y/\Delta x = 0/0$, por la cual no tenemos el menor interés. Sólo se producen el misterio y la confusión si seguimos el camino de Leibniz y de muchos de sus sucesores, y decimos algo como esto: «Δx no tiende a cero; por el contrario, el «último valor» de Δx no es cero, sino «una cantidad infinitamente pequeña», una «diferencial», llamada dx, y, análogamente, Δy tiene un «último» valor infinitamente pequeño, dy. El cociente verdadero de estas diferenciales infinitamente pequeñas es un número ordinario: $f'(x) = dy/dx$.» Por esó Leibniz llamó a la derivada *«cociente diferencial»*. Esas cantidades infinitamente pequeñas eran consideradas como una nueva clase de números, no iguales a cero, sino más pequeñas que cualquier número real. Sólo los que poseían una verdadera intuición matemática podían captar este concepto, por lo que se creía que el cálculo era genuinamente difícil, ya que no todos tienen o pueden desarrollar esa intuición. De igual manera, se consideraba la integral como una suma de infinitas «cantidades infinitamente pequeñas» $f(x)dx$, y se tenía la creencia de que dicha suma *era* la integral o área, considerándose como algo accesorio el cálculo de su valor mediante el *límite de una suma finita de números ordinarios*, $f(x_j)\Delta x$. Hoy damos de lado el deseo de una explicación «directa» y *definimos* la integral como límite de una suma finita. De esta manera, se eliminan todas las dificultades y se asienta sobre una base sólida todo lo que es fundamental en el cálculo.

A pesar de este posterior desarrollo, se retuvo la notación de Leibniz, dy/dx, para la derivada $f'(x)$, y $\int f(x)\, dx$ para la integral, pues ha demostrado ser extremadamente útil. No plantea ningún inconveniente si recordamos que son simples símbolos para representar el paso al límite. La notación de Leibniz ofrece la ventaja de que los límites de cocientes y de sumas pueden manejarse en ciertos aspectos «como si» realmente fueran cocientes o sumas. El poder sugestivo de esos símbolos ha inducido muchas veces a ciertas personas a atribuirles un significado que está por entero fuera de la matemática.

Si nos resistimos a esa tentación, la notación de Leibniz es, al menos, una excelente taquigrafía, que sustituye a la notación explícita más complicada del paso al límite. De hecho es indispensable en ciertos aspectos más avanzados de la teoría.

Por ejemplo, la regla *d*) de la página 439 para derivar la función inversa $x = g(y)$ de $y = f(x)$ dice que $g'(y)\, f'(y) = 1$. En la notación de Leibniz se tiene:

$$\frac{dx}{dy} \cdot \frac{dy}{dx} = 1,$$

«como si» las «diferenciales» pudieran reducirse en una expresión que parece ser un quebrado ordinario. Igualmente, la regla *e*) de la página 440, para la derivada de una función compuesta $z = k(x)$, siendo $z = g(y)$, e $y = f(x)$, puede expresarse así:

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

La notación de Leibniz tiene además la ventaja de fijar la atención en las *cantidades* x, y, z, más que en su conexión funcional explícita. Esta última expresa un *procedimiento*, una *operación*, que produce una cantidad y a partir de otra x; p. ej., la función $y = f(x) = x^2$ produce una cantidad y igual al cuadrado de la cantidad x. La operación (elevar al cuadrado) requiere la atención del matemático, pero el físico y el ingeniero se interesan en primer lugar por las propias cantidades. De ahí que el énfasis de la notación de Leibniz sobre las cantidades sea particularmente atractivo para los que se interesan por la matemática aplicada.

Podemos añadir otra observación. Mientras que «las diferenciales» se han descartado definitivamente y de forma poco honorable, en cuanto cantidades infinitamente pequeñas, la palabra *diferencial* ha entrado por la puerta falsa, pero para designar un concepto legítimo y útil. Significa ahora simplemente una diferencia Δx cuando Δx es pequeño en relación con las otras cantidades que intervienen en el cálculo. No resulta oportuno examinar aquí el valor de este concepto en el cálculo aproximado, ni tampoco podemos considerar otras nociones matemáticas perfectamente legítimas, que se designan con el mismo nombre de «diferencial», algunas de las cuales han demostrado ser sumamente útiles en el cálculo y en sus aplicaciones a la geometría.

V. EL TEOREMA FUNDAMENTAL DEL CALCULO

1. El teorema fundamental.—La noción de integral, y hasta cierto punto la de derivada, estaban bastante desarrolladas antes de los

trabajos de Newton y Leibniz. Para que se iniciara la prodigiosa evolución del nuevo análisis matemático sólo era necesario un descubrimiento muy sencillo. Los dos pasos al límite aparentemente inconexos en que se basan tanto la derivación como la integración de una función están íntimamente relacionados. De hecho son inversos entre sí, como la suma y la resta o la multiplicación y la división. No existen por separado cálculo diferencial y cálculo integral, sino que hay un solo *cálculo.*

El mérito enorme de Newton y Leibniz consiste en haber sido los primeros en reconocer y utilizar este *teorema fundamental del cálculo.* Naturalmente, su descubrimiento se encontraba en la trayectoria del pensamiento científico y resulta natural que distintas personas llegaran a comprender tal situación, independientemente y casi al mismo tiempo.

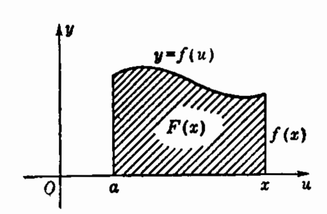


FIG. 274.—La integral como función del extremo superior.

Para formular el teorema fundamental, consideremos la integral de una función $y = f(x)$ desde el límite inferior fijo a al límite superior variable x. Para evitar la confusión entre el límite superior de integración x y la variable x que aparece en el símbolo $f(x)$, escribiremos esta integral de la manera siguiente (véase pág. 422):

$$F(x) = \int_a^x f(u)\, du, \qquad [1]$$

para indicar que deseamos estudiar la integral en cuanto función $F(x)$ del extremo superior x (Fig. 274). Esta función $F(x)$ es el área bajo la curva $y = f(u)$ desde el punto $u = a$ hasta el punto $u = x$. Algunas veces se llama integral «indefinida» a esta $F(x)$ que depende del extremo superior variable.

El teorema fundamental del cálculo dice lo siguiente:

La derivada de la integral indefinida [1] *como función de* x *es igual al valor de* f(u) *en el punto* x:

$$F'(x) = f(x).$$

En otras palabras, el proceso de integración que conduce de la función f(x) *a la* F(x) *queda anulado o invertido por el de derivación aplicado a* F(x).

Intuitivamente, la demostración es muy sencilla; depende de la interpretación de $F(x)$ como área y resulta oscura si se intenta representar $F(x)$ por una curva y la derivada $F'(x)$ mediante su pen-

diente. En lugar de esta interpretación geométrica inicial de la derivada, retengamos el significado geométrico de la integral $F(x)$ y procedamos analíticamente a derivar $F(x)$. La diferencia

$$F(x_1) - F(x)$$

es sencillamente el área entre x y x_1 en la figura 275, y vemos que está comprendida entre los valores $(x_1 - x)m$ y $(x_1 - x)M$; es decir:

$$(x_1 - x)m < F(x_1) - F(x) < (x_1 - x)M,$$

donde M y m representan, respectivamente, el máximo y el mínimo de $f(u)$ en el intervalo (x, x_1). En efecto, dichos productos son las

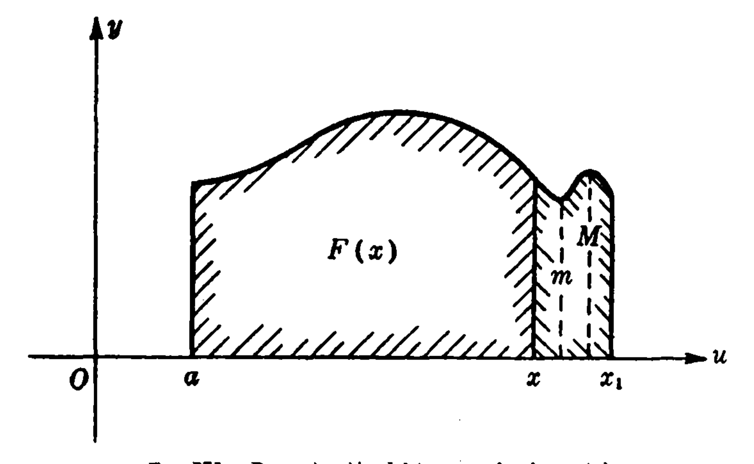


FIG. 275.—Demostración del teorema fundamental.

áreas de dos rectángulos tales que el uno incluye el área de la curva, mientras que el otro está incluído en ésta. En consecuencia:

$$m < \frac{F(x_1) - F(x)}{x_1 - x} < M.$$

Supondremos que la función $f(u)$ es continua, por lo que, cuando x_1 tiende a x, M y m tienden ambos a $f(x)$. De este hecho resulta:

$$F'(x) = \lim \frac{F(x_1) - F(x)}{x_1 - x} = f(x), \qquad [2]$$

como habíamos afirmado. Intuitivamente, esto expresa que la «velocidad» con que varía el área bajo la curva $y = f(x)$ cuando x crece, es igual a la altura de la curva en el punto x.

En algunos textos, este punto importante del teorema fundamental queda oscurecido por una nomenclatura mal elegida. Muchos autores introducen primero la derivada y definen después la «integral indefinida» como la inversa de la derivada, afirmando que $G(x)$ es una integral indefinida de $f(x)$, si

$$G'(x) = f(x).$$

Así, este procedimiento relaciona inmediatamente la derivación con la «integral», y después se introduce el concepto de «integral definida» como área o como límite de una suma, sin advertir explícitamente que en este caso la palabra «integral» significa ahora algo por completo distinto. De esta manera, se introduce subrepticiamente el punto fundamental de la teoría, lo que dificulta seriamente los esfuerzos del estudiante para alcanzar una comprensión clara de estas cosas. Resulta preferible denominar a las funciones $G(x)$, para las cuales $G'(x) = f(x)$, *funciones primitivas de* $f(x)$ y no «integrales indefinidas». En esta forma, el teorema fundamental puede enunciarse así:

La función F(x), *integral de* f(u), *cuyo extremo inferior es fijo y variable el superior* x, *es una función primitiva de* f(x).

Decimos «una» función primitiva y no «la» función primitiva, pues es evidente que si $G(x)$ es una función primitiva de $f(x)$, también lo será

$$H(x) = G(x) + c \quad \text{(siendo } c \text{ una constante arbitraria),}$$

puesto que $H'(x) = G'(x)$. También se verifica la propiedad recíproca. *Dos funciones primitivas* G(x) *y* H(x) *sólo pueden diferir en una constante.* Pues la diferencia $U(x) = G(x) - H(x)$ tiene la derivada $U'(x) = G'(x) - H'(x) = f(x) - f(x) = 0$, y, por tanto, es constante, ya que una función representada por una gráfica horizontal debe ser constante.

Esto nos conduce a una importante regla para determinar el valor de una integral entre a y b, siempre que conozcamos una función primitiva $G(x)$ de $f(x)$. De acuerdo con el teorema fundamental,

$$F(x) = \int_a^x f(u)\,du$$

es también una primitiva de $f(x)$. Por tanto, $F(x) = G(x) + c$, donde c es una constante que se determina recordando que $F(a) = \int_a^a f(u)du = 0$, conduce a la igualdad $0 = G(a) + c$; o sea, $c = -G(a)$. Así, la inte-

gral definida entre los límites a y x será $F(x) = \int_a^x f(u)du = G(x) - G(a)$, y si escribimos b en lugar de x:

$$\int_a^b f(u)\,du = G(b) - G(a), \qquad [3]$$

prescindiendo de cuál sea la particular función primitiva $G(x)$ elegida. En otras palabras:

Para calcular la integral definida $\int_a^b f(x)dx$ *sólo necesitamos determinar una función* $G(x)$ *tal que* $G'(x) = f(x)$, *formando entonces la diferencia* $G(b) - G(a)$.

2. Primeras aplicaciones. Integración de x^r, cos x, sen x, arc tg x. Es imposible dar aquí una idea adecuada del alcance del teorema fundamental, pero los ejemplos siguientes proporcionarán alguna indicación. Con mucha frecuencia, en los problemas reales que aparecen en mecánica, física o en matemática pura, se tiene que hallar el valor de una integral definida. El método directo de determinar la integral como límite de una suma puede resultar difícil. Por otra parte, según hemos visto anteriormente en III, es relativamente fácil efectuar cualquier derivación y acumular toda la información obtenida por este procedimiento, con lo que puede invertirse cada fórmula de derivación $G'(x) = f(x)$, proporcionando así una función primitiva $G(x)$ de $f(x)$. Mediante la fórmula [3], se puede calcular la integral de $f(x)$ entre dos límites cualesquiera.

Por ejemplo, si queremos encontrar la integral de x^2, x^3 o x^n, podemos proceder ahora de manera mucho más sencilla que anteriormente. Por la fórmula de derivación de x^n sabemos que la derivada de x^n es nx^{n-1}, por lo que la derivada de

$$G(x) = \frac{x^{n+1}}{n+1} \qquad (n \neq -1)$$

es:

$$G'(x) = \frac{n+1}{n+1}x^n = x^n.$$

En consecuencia, $x^{n+1}/(n+1)$ es una función primitiva de $f(x) = x^n$, y se tiene inmediatamente:

$$\int_a^b x^n\,dx = G(b) - G(a) = \frac{b^{n+1} - a^{n+1}}{n+1}$$

Este proceso es mucho más sencillo que el de hallar directamente la integral como límite de una suma.

Antes hemos establecido, con mayor generalidad, que para cualquier valor racional s, positivo o negativo, la función x^s tiene como derivada sx^{s-1}, y, por tanto, para $s = r + 1$, la función

$$G(x) = \frac{1}{r + 1} x^{r+1}$$

tiene la derivada $f(x) = G'(x) = x^r$. (Se supone $r \neq -1$; es decir, $s \neq 0$). En consecuencia, $x^{r+1}/r + 1$ es una función primitiva o «integral indefinida» de x^r, y se tiene (para a y b positivos, y $r \neq -1$):

$$\int_a^b x^r\,dx = \frac{1}{r + 1}(b^{r+1} - a^{r+1}). \qquad [4]$$

En [4] suponemos que en el intervalo de integración, el integrando x^r está definido y es continuo, lo que excluye $x = 0$, si $r < 0$. En consecuencia, en este caso se supone que a y b son positivos.

Para $G(x) = -\cos x$, tenemos que $G'(x) = \operatorname{sen} x$, de donde resulta:

$$\int_0^a \operatorname{sen} x\,dx = -(\cos a - \cos 0) = 1 - \cos a.$$

De forma análoga, puesto que para $G(x) = \operatorname{sen} x$ se tiene $G'(x) = \cos x$, se deduce:

$$\int_0^a \cos x\,dx = \operatorname{sen} a - \operatorname{sen} 0 = \operatorname{sen} a.$$

Se obtiene un resultado particularmente interesante de la fórmula de derivación de la función arco tangente, D arc tg $x = 1/(1 + x^2)$, de donde la función arc tg x es una primitiva de $1/(1 + x^2)$, y de la fórmula [3] se deduce el resultado siguiente:

$$\operatorname{arc\,tg} b - \operatorname{arc\,tg} 0 = \int_0^b \frac{1}{1 + x^2}\,dx.$$

Pero, arc tg $0 = 0$, ya que a un ángulo nulo corresponde también una tangente nula, por lo que podemos escribir:

$$\operatorname{arc\,tg} b = \int_0^b \frac{1}{1 + x^2}\,dx. \qquad [5]$$

En particular, si $b = 1$, arc tg b será igual a $\pi/4$, puesto que el

valor 1 de la tangente corresponde a un ángulo de 45º, o $\pi/4$ si se mide en radianes. Así, se obtiene la notable fórmula:

$$\pi/4 = \int_0^1 \frac{1}{1+x^2}\,dx. \qquad [6]$$

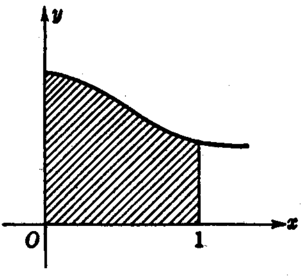


FIG. 276.—El área, desde 0 a 1, bajo la curva $y = 1/(1 + x^2)$ es $\pi/4$.

Esto demuestra que el área bajo la curva de la función $y = 1/(1 + x^2)$, desde $x = 0$ hasta $x = 1$, es igual a un cuarto de la de un círculo de radio 1.

3. La fórmula de Leibniz para π. Este último resultado conduce a uno de los más bellos descubrimientos matemáticos del siglo XVII: la serie alternada de Leibniz para π.

$$\frac{\pi}{4} = \frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \dots \qquad [7]$$

Mediante el símbolo $+ \dots$, queremos expresar que la sucesión de las «sumas parciales» finitas, formadas al interrumpir la expresión a la derecha después de tomar n términos, converge hacia $\pi/4$ cuando n crece indefinidamente.

Para demostrar esta famosa fórmula, basta recordar la progresión geométrica finita:

$$\frac{1-q^n}{1-q} = 1 + q + q^2 + \cdots + q^{n-1},$$

o bien:

$$\frac{1}{1-q} = 1 + q + q^2 + \cdots + q^{n-1} + \frac{q^n}{1-q}$$

En esta identidad algebraica hacemos $q = -x^2$ y obtenemos:

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots + (-1)^{n-1}x^{2n-2} + R_n, \qquad [8]$$

donde el «resto» R_n es:

$$R_n = (-1)^n \frac{x^{2n}}{1+x^2}$$

Ahora puede integrarse la ecuación [8] entre los límites 0 y 1; de acuerdo con la regla *a)* de la página 437, deberemos escribir a la derecha la suma de las integrales de los sucesivos términos, y apli-

cando la fórmula [4] resulta: $\int_a^b x^m dx = (b^{m+1} - a^{m+1})/(m+1)$; o sea, $\int_0^1 x^m dx = 1/(m+1)$, y finalmente,

$$\int_0^1 \frac{dx}{1+x^2} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots + (-1)^{n-1}\frac{1}{2n-1} + T_n, \quad [9]$$

siendo $T_n = (-1)^n \int_0^1 \frac{x^{2n}}{1+x^2} dx.$

Según [5], el segundo miembro de [9] es igual a $\pi/4$, y la diferencia entre $\pi/4$ y la suma parcial

$$S_n = 1 - \frac{1}{3} + \frac{1}{5} + \cdots + \frac{(-1)^{n-1}}{2n-1}$$

es $\pi/4 - S_n = T_n$. Queda por demostrar que T_n tiende a cero al crecer n. Pero, por ser

$$\frac{x^{2n}}{1+x^2} \leqslant x^{2n} \qquad \text{(para } 0 \leqslant x \leqslant 1\text{).}$$

si recordamos la fórmula [13] de I, según la cual, $\int_a^b f(x)dx \leqslant \int_a^b g(x)dx$ si $f(x) \leqslant g(x)$ y $a < b$, vemos que

$$|T_n| = \int_0^1 \frac{x^{2n}}{1+x^2} dx \leqslant \int_0^1 x^{2n} dx;$$

y como el segundo miembro es igual a $1/(2n+1)$, según hemos comprobado antes (fórmula [4]), vemos que $|T_n| < 1/(2n+1)$. Por tanto,

$$\left|\frac{\pi}{4} - S_n\right| < \frac{1}{2n+1}$$

Esto demuestra que S_n tiende a $\pi/4$ al aumentar n, ya que $1/(2n+1)$ tiende a cero. Queda así demostrada la fórmula de Leibniz.

VI. LAS FUNCIONES EXPONENCIAL Y LOGARÍTMICA

Los conceptos fundamentales del cálculo proporcionan una teoría más adecuada de las funciones logarítmica y exponencial que el método «elemental» utilizado en la enseñanza media. Allí se empieza generalmente por las potencias enteras a^n de un número positivo a, y se define después $a^{1/m} = \sqrt[m]{a}$, obteniendo así el valor de a^r para

cualquier número racional $r = n/m$. A continuación se define el valor de a^x para cualquier número irracional x de modo que a^x sea una función continua de x, punto delicado, que se omite en la enseñanza elemental. Finalmente, se define el logaritmo de y en base a,

$$x = \log_a y,$$

como la función inversa de $y = a^x$.

En la teoría que sigue de estas funciones, basada en el cálculo, se invierte este orden. Empezamos por el logaritmo y obtenemos después la función exponencial.

1. Definición y propiedades del logaritmo. El número «e» de Euler. Definiremos el logaritmo, o, más concretamente, el «logaritmo natural» $F(x) = \log x$ [dejamos para más adelante (pág. 457) su relación con el logaritmo de base 10], como el área comprendida bajo la curva $y = 1/u$ desde $u = 1$ hasta $u = x$, o, lo que significa lo mismo, mediante la integral

$$F(x) = \log x = \int_1^x \frac{1}{u}\, du \qquad [1]$$

(véase Fig. 5). La variable x puede ser un número positivo cualquiera. Queda excluído el cero, ya que el integrando tiende a infinito cuando x tiende a cero.

Resulta natural que estudiemos la función $F(x)$, pues sabemos que la primitiva de cualquier potencia x^n es una función $x^{n+1}/(n + 1)$ del mismo tipo, excepto para $n = 1$. En este último caso, el denominador $n + 1$ se anula y la fórmula [4] de la página 450 carece de significado. Así cabe esperar que la integración de $1/x$ o de $1/u$ conduzca a algún tipo nuevo—e interesante—de función.

Aunque consideramos [1] como definición de la función $\log x$, no podremos decir que la *conocemos* hasta que hayamos deducido sus propiedades y encontrado medios para calcularla numéricamente. Es éste un ejemplo típico de los métodos modernos de la matemática, consistentes en empezar por conceptos generales, tales como área e integral, estableciendo seguidamente definiciones, como la [1], que se basan en ellos, deduciendo después las propiedades de los entes así definidos y obteniendo finalmente expresiones explícitas para su cálculo numérico.

La primera propiedad importante de $\log x$ es una consecuencia inmediata del teorema fundamental de V, el cual proporciona la ecuación

$$F'(x) = 1/x. \qquad [2]$$

De [2] se deduce que la derivada es siempre positiva, lo que confirma un hecho evidente: que la función $\log x$ es monótona creciente para valores crecientes de x.

La principal propiedad del logaritmo queda expresada por la fórmula

$$\log a + \log b = \log (ab). \qquad [3]$$

Es bien conocida la importancia de esta fórmula en la aplicación práctica de los logaritmos al cálculo numérico. Intuitivamente, la fórmula [3] se podría obtener considerando las áreas que definen las tres cantidades $\log a$, $\log b$, y $\log (ab)$. Pero preferimos deducirla mediante un razonamiento típico del cálculo. Junto con la función $F(x) = \log x$, consideremos la siguiente:

$$k(x) = \log (ax) = \log w = F(w),$$

donde $w = f(x) = ax$, siendo a una constante positiva cualquiera. Es fácil derivar $k(x)$ mediante la regla de la página 441: $k'(x) = F'(w)f'(x)$. De [2], y puesto que $f'(x) = a$, resulta:

$$k'(x) = a/w = a/ax = 1/x.$$

En consecuencia, $k(x)$ tiene la misma derivada que $F(x)$, y de acuerdo con lo dicho en la página 448, tendremos:

$$\log (ax) = k(x) = F(x) + c,$$

donde c es una constante que no depende del valor particular de x, y que se determina sin más que sustituir x por el número 1. De la definición [1], se deduce que

$$F(1) = \log 1 = 0,$$

puesto que la integral definida tiene iguales ambos extremos para $x = 1$. Se obtiene así:

$$k(1) = \log (a \cdot 1) = \log a = \log 1 + c = c,$$

lo que proporciona $c = \log a$; o sea, que para todo x se verifica

$$\log (ax) = \log a + \log x. \qquad [3a]$$

Haciendo $x = b$, se obtiene la fórmula buscada [3].

En particular, para $a = x$, encontramos sucesivamente:

$$\log (x^2) = 2 \log x,$$
$$\log (x^3) = 3 \log x,$$
$$\cdots\cdots$$
$$\log (x^n) = n \log x. \qquad [4]$$

La ecuación [4] prueba que al aumentar x los valores de log x tienden a infinito, pues el logaritmo es una función monótona creciente, y se tiene, p. ej.,

$$\log (2^n) = n \log 2,$$

que tiende a infinito con n. Además,

$$0 = \log 1 = \log \left(x \cdot \frac{1}{x}\right) = \log x + \log \frac{1}{x};$$

de forma que

$$\log \frac{1}{x} = -\log x. \qquad [5]$$

Finalmente,

$$\log x^r = r \log x \qquad [6]$$

para cualquier número racional $r = m/n$. En efecto, si se hace $x^r = u$, se tiene:

$$n \log u = \log u^n = \log x^{\frac{m}{n} \cdot n} = \log x^m = m \log x;$$

de donde,

$$\log x^{\frac{m}{n}} = \frac{m}{n} \log x.$$

Dado que log x es una función continua y monótona de x, que toma el valor 0 para $x = 1$, y tiende a infinito con x, debe existir un número mayor que 1, tal que para él se tenga $\log x = 1$.

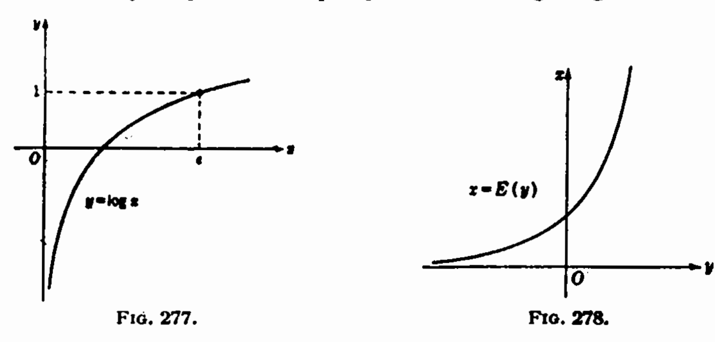


FIG. 277. FIG. 278.

Siguiendo el ejemplo de Euler, llamaremos e a este número (más adelante se demostrará la equivalencia con la definición de la página 309). Así, pues, e está definido por la igualdad

$$\log e = 1. \qquad [7]$$

Hemos introducido el número e mediante una propiedad intrínseca que asegura su *existencia*. Ahora proseguiremos nuestro análisis, del que se deducirán como *consecuencia* fórmulas explícitas que proporcionan aproximaciones tan exactas como se desee para el cálculo numérico de e.

2. La función exponencial.—Resumiendo los resultados anteriores, vemos que la función $F(x) = \log x$ tiene el valor cero para $x = 1$; crece monótonamente hasta infinito, pero con pendiente $1/x$ decreciente, y para valores positivos de x menores que 1, está dada por $\log 1/x$ tomado con signo negativo, por lo que $\log x$ tiende a $-\infty$ cuando $x \to 0$.

Debido al carácter de monotonía de $y = \log x$, podemos considerar la función inversa

$$x = E(y),$$

cuya gráfica (Fig. 278) se obtiene de la forma usual a partir de la de $y = \log x$ (Fig. 277); está definida para todo valor de y entre $-\infty$ y $+\infty$. Cuando y tiende a $-\infty$, $E(y)$ tiende a cero, y cuando y tiende a $+\infty$, $E(y)$ tiende a $+\infty$.

La función E tiene la siguiente propiedad fundamental:

$$E(a) \cdot E(b) = E(a + b) \qquad [8]$$

para cualquier par de valores a y b. Esta igualdad es sólo otra manera de expresar la ley [3] del logaritmo. Pues si es $E(b) = x$, $E(a) = z$ (es decir, $b = \log x$; $a = \log z$), se tiene:

$$\log xz = \log x + \log z = b + a,$$

y, en consecuencia, resulta:

$$E(b + a) = xz = E(a) \cdot E(b),$$

como se quería demostrar. Ya que, por definición, $\log e = 1$, tenemos

$$E(1) = e,$$

de donde se sigue, teniendo en cuenta [8], que $e^2 = E(1)E(1) = E(2)$, etcétera. En general,

$$E(n) = e^n$$

para cualquier entero n. De forma análoga, $E(1/n) = e^{1/n}$, de modo que $E(p/q) = E(1/q) \cdots E(1/q) = [e^{1/q}]^p$; por tanto, haciendo $p/q = r$, se tiene, finalmente,

$$E(r) = e^r$$

para cualquier valor racional r. Resulta así apropiado *definir* la operación de elevar el número e a una potencia irracional escribiendo

$$e^y = E(y)$$

para cualquier número real y, puesto que la función E es continua para todos los valores de y, e idéntica al valor de e^y para y racional. Ahora podemos expresar la ley fundamental [8] de la función E o *función exponencial*, como también se llama, mediante la ecuación

$$e^a \cdot e^b = e^{a+b}, \qquad [9]$$

que queda así demostrada para números racionales o irracionales arbitrarios a y b.

En toda esta discusión acerca de las funciones exponencial y logarítmica, nos hemos referido al número e, como la «base» o «base natural» de los logaritmos. Es fácil pasar de la base e a cualquier otro número positivo. Comencemos por considerar el logaritmo (natural)

$$\alpha = \log a,$$

de donde se deduce que

$$a = e^\alpha = e^{\log a}.$$

Definamos ahora a^x mediante la expresión compuesta

$$z = a^x = e^{\alpha x} = e^{x \log a}. \qquad [10]$$

Por ejemplo,

$$10^x = e^{x \log 10}.$$

Llamamos *logaritmo de base a* a la función inversa de a^x, y vemos en seguida que el *logaritmo natural* de z es x veces α; en otras palabras, el logaritmo de un número z en base a resulta de dividir el logaritmo natural de z por el logaritmo natural de a, que es una constante. Para $a = 10$, este número, con cuatro cifras exactas, es

$$\log 10 = 2{,}303.$$

3. Fórmulas de derivación de e^x, a^x, x^s.—Puesto que hemos definido la función exponencial $E(y)$ como la inversa de $y = \log x$, de la regla para derivar funciones inversas (véase III) resulta:

$$E'(y) = \frac{dx}{dy} = \frac{1}{\dfrac{dy}{dx}} = \frac{1}{1/x} = x = E(y),$$

es decir: *La función exponencial natural es idéntica a su derivada*:

$$E'(y) = E(y). \qquad [11]$$

Éste es, en realidad, el origen de todas las propiedades de la función exponencial y la razón fundamental de su importancia en matemática aplicada, según se verá en los párrafos siguientes. Utilizando la notación de Leibniz, podemos escribir [11] de la manera siguiente:

$$\frac{d}{dx} e^x = e^x. \qquad [11a]$$

Con mayor generalidad, derivando la función compuesta: $f(x) = e^{\alpha x}$, se tiene, aplicando una conocida fórmula,

$$f'(x) = \alpha e^{\alpha x} = \alpha f(x).$$

De donde, para $\alpha = \log a$, encontramos que la función

$$f(x) = a^x$$

tiene la derivada

$$f'(x) = a^x \log a.$$

Podemos definir ahora la función

$$f(x) = x^s$$

para cualquier exponente real s y variable positiva x, escribiendo

$$x^s = e^{s \log x}.$$

Aplicando nuevamente la regla de derivación de las funciones compuestas, $f(x) = e^{sz}$, $z = \log x$, encontramos $f'(x) = se^{sz}(1/x) = sx^s(1/x)$, y, en consecuencia,

$$f'(x) = sx^{s-1},$$

de acuerdo con un resultado anterior para s racional.

4. Expresiones explícitas de e, e^x y log x, en forma de límite. Con el fin de hallar fórmulas explícitas para estas funciones, utilizaremos las reglas de derivación de las funciones exponencial y logarítmica. Dado que la derivada de la función $\log x$ es $1/x$, por medio de la definición de derivada obtenemos la relación

$$\frac{1}{x} = \lim \frac{\log x_1 - \log x}{x_1 - x} \qquad \text{cuando} \qquad x_1 \to x.$$

Si escribimos $x_1 = x + h$ y hacemos tender h a cero siguiendo los valores de la sucesión

$$h = 1/2, 1/3, 1/4, \dots, 1/n, \dots,$$

resulta, aplicando las reglas del cálculo logarítmico,

$$\frac{\log\left(x + \frac{1}{n}\right) - \log x}{1/n} = n \log \frac{x + \frac{1}{n}}{x} = \log\left[\left(1 + \frac{1}{nx}\right)^n\right] \to \frac{1}{x}$$

Si se pone $z = 1/x$ y se hace uso nuevamente de las mismas reglas, se encuentra:

$$z = \text{lím} \log\left[\left(1 + \frac{z}{n}\right)^n\right] \quad \text{cuando} \quad n \to \infty;$$

o sea, mediante la función exponencial,

$$e^z = \text{lím}\left(1 + \frac{z}{n}\right)^n \quad \text{cuando} \quad n \to \infty. \qquad [12]$$

He aquí la famosa fórmula que define la función exponencial en forma de límite. En particular, para $z = 1$, se tiene:

$$e = \text{lím}\,(1 + 1/n)^n, \qquad [13]$$

y para $z = -1$,

$$\frac{1}{e} = \text{lím}\,(1 - 1/n)^n. \qquad [13a]$$

Estas expresiones conducen inmediatamente a desarrollos en forma de serie. Mediante el teorema del binomio, se tiene:

$$\left(1 + \frac{x}{n}\right)^n = 1 + n\frac{x}{n} + \frac{n(n-1)}{2!}\frac{x^2}{n^2} + \frac{n(n-1)(n-2)}{3!}\frac{x^3}{n^3} + \cdots + \frac{x^n}{n^n};$$

o sea,

$$\left(1 + \frac{x}{n}\right)^n = 1 + \frac{x}{1!} + \frac{x^2}{2!}\left(1 - \frac{1}{n}\right) + \frac{x^3}{3!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) + \cdots +$$
$$+ \frac{x^n}{n!}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\cdots\left(1 - \frac{n-2}{n}\right)\left(1 - \frac{n-1}{n}\right).$$

Resulta conveniente, y no supone gran dificultad, justificar completamente (se omiten aquí los detalles) que es posible efectuar el paso

al límite para $n \to \infty$, reemplazando en cada término $1/n$ por 0. Esto proporciona la famosa serie para e^x:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots, \qquad [14]$$

y, en particular, se obtiene para e el desarrollo

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \cdots,$$

que establece la identidad de e con el número definido en la página 309. Para $x = -1$, se tiene la serie

$$\frac{1}{e} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} + \cdots,$$

la cual proporciona una excelente aproximación numérica con sólo muy pocos términos, siendo el error total introducido al considerar únicamente los n primeros términos inferior al valor absoluto del término de lugar $n + 1$.

Si utilizamos la fórmula de derivación de la función exponencial, puede obtenerse una expresión interesante para el logaritmo. Se tiene

$$\text{lím}\,\frac{e^h - 1}{h} = \text{lím}\,\frac{e^h - e^0}{h} = 1$$

cuando $h \to 0$, debido a que este límite es la derivada de e^y para $y = 0$, y ésta es igual a $e^0 = 1$. Sustituyamos en esta fórmula h por los valores z/n, donde z es un número arbitrario, y n toma los sucesivos valores enteros. Obtenemos así

$$n\,\frac{e^{z/n} - 1}{z} \to 1,$$

o

$$n(\sqrt[n]{e^z} - 1) \to z$$

al tender n a infinito. Si hacemos $z = \log x$ o $e^z = x$, se obtiene finalmente

$$\log x = \text{lím}\, n(\sqrt[n]{x} - 1) \quad \text{cuando} \quad n \to \infty. \qquad [15]$$

Como $\sqrt[n]{x} \to 1$ cuando $n \to \infty$ (véase pág. 334), [15] representa el logaritmo como límite de un producto, uno de cuyos factores tiende a cero y el otro a infinito.

Ejemplos y ejercicios varios.—Al incluir las funciones exponencial y logarítmica podemos considerar ahora una amplia clase de funciones y tener acceso a numerosas aplicaciones.

Obténganse las derivadas de:

1. $x(\log x - 1)$.
2. $\log(\log x)$.
3. $\log(x + \sqrt{1 + x^2})$.
4. $\log(x + \sqrt{1 - x^2})$.
5. e^{-x^2}.
6. e^{e^x} (función compuesta e^z, con $z = e^x$).
7. x^x (Hágase $x^x = e^{x \log x}$).
8. $\log \operatorname{tg} x$.
9. $\log \operatorname{sen} x$; $\log \cos x$.
10. $x/\log x$.

Determínense los máximos y mínimos de las funciones:

11. xe^{-x}.
12. x^2e^{-x}.
13. xe^{-ax}.

*14. Determínese el lugar geométrico de los máximos de la curva $y = xe^{-ax}$, al variar a.

15. Demuéstrese que las sucesivas derivadas de la función e^{-x^2} están formadas por el producto de la función e^{-x^2} y un polinomio en x.

*16. Demuéstrese que la n-ésima derivada de e^{-1/x^2} es igual a esta misma función, multiplicada por un polinomio de grado $2n - 2$, dividido por x^{3n}.

*17. *Derivación logarítmica.* Utilizando la propiedad fundamental del logaritmo, puede efectuarse a menudo la derivación de un producto de manera más sencilla. Para un producto de la forma:

$$p(x) = f_1(x)f_2(x) \cdots f_n(x),$$
$$D(\log p(x)) = D(\log f_1(x)) + D(\log f_2(x)) + \cdots + D(\log f_n(x)),$$

de donde, por la regla de derivación de las funciones compuestas, resulta:

$$\frac{p'(x)}{p(x)} = \frac{f_1'(x)}{f_1(x)} + \frac{f_2'(x)}{f_2(x)} + \cdots + \frac{f_n'(x)}{f_n(x)}$$

Utilícese este procedimiento para derivar

a) $x(x + 1)(x + 2) \cdots (x + n)$; b) xe^{-ax^2}.

5. Serie logarítmica. Cálculo numérico.—La fórmula [15] no es adecuada para el cálculo numérico de los logaritmos. Una expresión explícita distinta y más útil, de gran importancia teórica, se presta mejor para este propósito. La obtendremos siguiendo el método utilizado en la página 451 para calcular π, a partir de la definición del logaritmo por la fórmula [1]. Se necesita un pequeño paso previo; en vez de operar con $\log x$, intentaremos expresar $y = \log(1 + x)$, compuesta por las funciones $y = \log z$ y $z = x + 1$. Tenemos así $\frac{dy}{dx} = \frac{dy}{dz} \cdot \frac{dz}{dx} = \frac{1}{z} \cdot 1 = \frac{1}{1 + x}$, de donde resulta que $\log(1 + x)$ es una función primitiva de $1/(1 + x)$, y de ello deducimos, por el teo-

rema fundamental, que la integral de $1/(1+u)$, desde 0 a x, es igual a $\log(1+x) - \log 1 = \log(1+x)$; simbólicamente:

$$\log(1+x) = \int_0^x \frac{1}{1+u}\,du. \qquad [16]$$

(Naturalmente, esta fórmula puede deducirse intuitivamente de la interpretación geométrica del logaritmo como área. Véase pág. 422.)

En la fórmula [16] introducimos, como hicimos en la página 452, la serie geométrica para $(1+u)^{-1}$, escribiendo

$$\frac{1}{1+u} = 1 - u + u^2 - u^3 + \cdots + (-1)^{n-1}u^{n-1} + (-1)^n \frac{u^n}{1+u},$$

donde, por precaución, preferimos no escribir una serie indefinida, sino compuesta de un número de términos, más un resto

$$R_n = (-1)^n \frac{u^n}{1+u}$$

Al introducir esta serie en [16] hacemos uso de la regla de que una suma (finita) de ese tipo puede integrarse término a término. La integral de u^s desde 0 hasta x es $x^{s+1}/(s+1)$, y se obtiene inmediatamente

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots + (-1)^{n-1}\frac{x^n}{n} + T_n,$$

donde el resto T_n está dado por

$$T_n = (-1)^n \int_0^x \frac{u^n}{1+u}\,du.$$

Demostremos ahora que T_n tiende a cero al aumentar n, siempre que x sea mayor que -1 y no exceda a $+1$; en otras palabras, para

$$-1 < x \leqslant 1,$$

debiendo observarse que se incluye el valor $x = +1$, pero queda excluído el $x = -1$. De acuerdo con nuestra hipótesis, en el intervalo de integración, u es mayor que un número $-\alpha$, que puede ser próximo a -1, pero que, en todo caso, es mayor que él, por lo que se tiene $0 < 1 - \alpha < 1 + u$. De ahí que en el intervalo de 0 a x, tengamos:

$$\left|\frac{u^n}{1+u}\right| \leqslant \frac{|u|^n}{1-\alpha},$$

y, por tanto,

$$|T_n| \leqslant \frac{1}{1-\alpha}\left|\int_0^x u^n\,du\right|,$$

o sea,

$$|T_n| \leqslant \frac{1}{1-\alpha}\,\frac{|x|^{n+1}}{n+1} \leqslant \frac{1}{1-\alpha}\,\frac{1}{n+1}$$

Puesto que $1-\alpha$ es constante, vemos que al tender n a infinito esta expresión tiende a cero, de forma que de

$$\left|\log(1+x) - \left\{x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots + (-1)^n\frac{x^n}{n}\right\}\right| \leqslant \frac{1}{1-\alpha}\,\frac{1}{n+1}, \qquad [17]$$

se obtiene la serie

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots, \qquad [18]$$

que es válida para $-1 < x \leqslant 1$.

En particular, para $x = 1$, resulta el curioso resultado:

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots, \qquad [19]$$

fórmula con una estructura similar a la obtenida para $\pi/4$.

La serie [18] no constituye un medio práctico para calcular los valores numéricos de los logaritmos, puesto que su intervalo de validez está reducido a los valores de $1+x$ comprendidos entre 0 y 2, y debido también a que su convergencia es tan lenta que es necesario tomar muchos términos para obtener un resultado medianamente aproximado. Puede obtenerse una expresión más adecuada por medio del siguiente artificio. Reemplazando x por $-x$ en [18], encontramos:

$$\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots \qquad [20]$$

Restando [20] de [18] y mediante la fórmula $\log a - \log b = \log a + \log(1/b) = \log(a/b)$, se tiene:

$$\log\frac{1+x}{1-x} = 2\left(x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots\right). \qquad [21]$$

No sólo esta serie converge mucho más de prisa, sino que además el primer miembro permite expresar el logaritmo de cualquier número positivo z, ya que $(1+x)/(1-x) = z$ tiene siempre una solu-

ción comprendida entre -1 y $+1$. Así, si queremos calcular log 3, hacemos $x = {}^1/_2$ y obtenemos

$$\log 3 = \log \frac{1 + {}^1/_2}{1 - {}^1/_2} = 2\left(\frac{1}{1 \cdot 2} + \frac{1}{3 \cdot 2^3} + \frac{1}{5 \cdot 2^5} + \cdots\right).$$

Con sólo seis términos, hasta el $\dfrac{2}{11 \cdot 2^{11}} = \dfrac{1}{11\,264}$, encontramos el valor $\log 3 = 1{,}0986$, que tiene cinco cifras exactas.

VII. ECUACIONES DIFERENCIALES

1. **Definición.**—El papel predominante que las funciones exponencial y trigonométricas desempeñan en el análisis matemático y en sus aplicaciones a la física, radica en que dichas funciones resuelven los tipos más sencillos de «ecuaciones diferenciales».

Una ecuación diferencial relaciona una función desconocida $u=f(x)$ con su derivada $u' = f'(x)$; es una ecuación que contiene u, u' y además, posiblemente, la variable independiente x; p. ej.,

$$u' = u + \text{sen}\,(xu)$$

o

$$u' + 3u = x^2.$$

Más en general, una ecuación diferencial puede contener la segunda derivada $u'' = f''(x)$ —e incluso derivadas de orden superior—, como, p. ej.,

$$u'' + 2u' + 3u = 0.$$

En cualquier caso, el problema consiste en encontrar una función $u = f(x)$ que satisfaga a la ecuación dada. Resolver una ecuación diferencial representa una amplia generalización del problema de la integración, en el sentido de encontrar la función primitiva de una función dada $g(x)$, lo que equivale a resolver la sencilla ecuación diferencial

$$u' = g(x).$$

Por ejemplo, las soluciones de la ecuación diferencial

$$u' = x^2$$

son las funciones $u = x^3/3 + c$, donde c es una constante arbitraria.

2. **La ecuación diferencial de la función exponencial. La desintegración radiactiva. La ley del crecimiento. Interés compuesto.**—La ecuación diferencial

$$u' = u \qquad [1]$$

tiene como solución la función exponencial $u = e^x$, ya que ésta es igual a su propia derivada. Con mayor generalidad, la función $u = ce^x$, donde c es una constante cualquiera, es una solución de [1]. Análogamente, la función

$$u = ce^{kx}, \qquad [2]$$

donde c y k son constantes cualesquiera, es una solución de la ecuación diferencial

$$u' = ku. \qquad [3]$$

Recíprocamente, cualquier función $u = f(x)$ que satisfaga a la ecuación [3] debe ser de la forma ce^{kx}. Pues si $x = h(u)$ es la función inversa de $u = f(x)$, se tiene:

$$h' = \frac{1}{u'} = \frac{1}{ku}$$

Pero $\dfrac{\log u}{k}$ es una función primitiva de $1/ku$, de forma que $x = h(u) = \dfrac{\log u}{k} + b$, siendo b una constante. De donde se deduce que

$$\log u = kx - bk,$$

y

$$u = e^{kx} \cdot e^{-bk}.$$

Haciendo e^{-bk} (que es una constante) igual a c, se tiene:

$$u = ce^{kx},$$

como queríamos demostrar.

La gran importancia de la ecuación diferencial [3] consiste en que gobierna numerosos procesos físicos, en los cuales una cantidad, u, de cierta sustancia es una función, $u = f(t)$, del tiempo, t, de tal forma que la cantidad u varía en cada instante proporcionalmente al valor de u en ese instante. En este caso, la *velocidad* en el instante t,

$$u' = f'(t) = \lim \frac{f(t_1) - f(t)}{t_1 - t},$$

es igual a ku, siendo k una constante, positiva si u aumenta y negativa si u disminuye. En uno y otro caso, u satisface a la ecuación diferencial [3]; por tanto,

$$u = ce^{kt}.$$

La constante c está determinada si conocemos la cantidad u_0 que existía en el instante $t = 0$. Obtendremos dicha cantidad si hacemos $t = 0$,

$$u_0 = ce^0 = c;$$

por tanto,

$$u = u_0 e^{kt}. \qquad [4]$$

Obsérvese que suponemos conocida la *velocidad de variación* de u, de donde deducimos la ley [4] que proporciona la *cantidad* de u existente en cualquier instante t. Éste es exactamente el problema recíproco de hallar la derivada de una función.

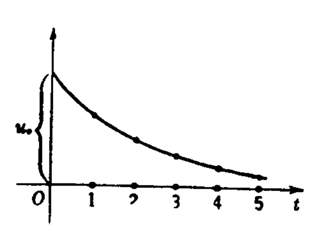


FIG. 279.—Decrecimiento exponencial, $u = u_0 e^{kt}$, $k < 0$.

Un ejemplo típico es el de la desintegración radiactiva. Sea $u = f(t)$ la cantidad de cierta sustancia radiactiva en el tiempo t; si suponemos que cada partícula individual de la sustancia tiene una determinada probabilidad de desintegrarse en un tiempo dado, y que esta probabilidad no depende de la presencia de las otras partículas, la velocidad a que se desintegra u en un instante dado, t, será proporcional a u, es decir, a la cantidad total de sustancia existente en ese momento. Por tanto, u satisface a [3] con una constante, k, negativa, que mide la velocidad del proceso de desintegración; esto es,

$$u = u_0 e^{kt}$$

Se sigue de ello que la fracción de u que se desintegra en dos intervalos iguales de tiempo es la misma, pues si u_1 es la cantidad que existe en el instante t_1 y u_2 la correspondiente a un instante posterior, t_2, se tiene:

$$\frac{u_2}{u_1} = \frac{u_0 e^{kt_2}}{u_0 e^{kt_1}} = e^{k(t_2 - t_1)},$$

que depende exclusivamente de la diferencia $t_2 - t_1$. Si deseamos hallar el tiempo que debe transcurrir para que se desintegre la mitad de la cantidad de sustancia inicialmente dada, deberemos determinar $s = t_2 - t_1$, de tal modo que

$$\frac{u_2}{u_1} = {}^1/_2 = e^{ks},$$

de donde resulta

$$ks = \log {}^1/_2, \qquad s = (-\log 2)/k, \qquad \text{o} \qquad k = (-\log 2)/s. \qquad [5]$$

Para cualquier sustancia radiactiva, el valor de s es el semiperíodo de desintegración; puede determinarse experimentalmente s o cualquier otro valor análogo (como, p. ej., el valor de r para el cual $u_2/u_1 = 999/1000$). Para el radio, el semiperíodo vale aproximadamente 1550 años, siendo $k = (\log {}^1/_2)/1550 = -0{,}0000447$. De aquí se deduce que

$$u = u_0 e^{-0{,}0000447t}$$

El interés compuesto nos ofrece otro ejemplo de una ley de crecimiento que es aproximadamente exponencial. Cierta cantidad de dinero, u_0 pesetas, se coloca al 3 % de interés compuesto, debiendo capitalizarse anualmente. Después de un año, el capital más los intereses será igual a

$$u_1 = u_0(1 + 0{,}03);$$

al cabo de dos años,

$$u_2 = u_1(1 + 0{,}03) = u_0(1 + 0{,}03)^2,$$

y transcurridos t años se tendrá

$$u_t = u_0(1 + 0{,}03)^t. \qquad [6]$$

Ahora bien: si en lugar de capitalizar por intervalos anuales se capitaliza cada mes, o cada n-ésima parte del año, después de t años, el capital y los intereses ascenderán a

$$u_0\left(1 + \frac{0{,}03}{n}\right)^{nt} = u_0\left[\left(1 + \frac{0{,}03}{n}\right)^n\right]^t.$$

Si n se hace muy grande, de tal modo que el interés se acumula al capital cada día o cada hora, al tender n a infinito, la cantidad entre corchetes, de acuerdo con VI, tiende a $e^{0{,}03}$, y en el límite, después de t años, el capital formado será igual a

$$u_0 \cdot e^{0{,}03t}, \qquad [7]$$

que corresponde a un proceso de interés compuesto continuo. También podemos calcular el tiempo s necesario para que se duplique un capital impuesto al 3 % anual de interés continuo. Tenemos $\dfrac{u_0 \cdot e^{0{,}03s}}{u_0} = 2$, por lo que $s = \dfrac{100}{3} \log 2 = 23{,}10$; el capital se duplicará al cabo de 23 años.

En lugar de proceder por etapas y efectuar después el paso al límite, pudimos haber obtenido la fórmula [7] diciendo simplemente que la velocidad de crecimiento u' del capital es proporcional a u, con el factor de proporcionalidad $k = 0{,}03$, por lo que

$$u' = ku, \quad (k = 0{,}03),$$

deduciendo entonces la fórmula [7] del resultado general [4].

3. Otros ejemplos. Movimiento vibratorio.—La función exponencial se presenta a menudo en combinaciones más complicadas; p. ej., la función

$$u = e^{-kx^2}, \quad [8]$$

en la que k es una constante positiva, es una solución de la ecuación diferencial

$$u' = -2kxu.$$

La función [8] es de importancia fundamental en probabilidades y estadística, ya que define la distribución «normal» de frecuencias.

Las funciones trigonométricas $u = \cos t$, $v = \operatorname{sen} t$, satisfacen también a ecuaciones diferenciales muy sencillas. Se tiene, en primer lugar.

$$u' = -\operatorname{sen} t = -v,$$
$$v' = \cos t = u,$$

que es un «sistema de dos ecuaciones diferenciales con dos funciones». Derivando nuevamente, se obtiene:

$$u'' = -v' = -u,$$
$$v'' = u' = -v,$$

por lo que ambas funciones u y v de la variable temporal t pueden considerarse como soluciones de la misma ecuación diferencial

$$z'' + z = 0, \quad [9]$$

la cual es una ecuación diferencial sencilla de «segundo orden»; es decir, que contiene la derivada segunda de z. Esta ecuación y su generalización con una constante positiva k^2,

$$z'' + k^2 z = 0, \quad [10]$$

de la cual son soluciones $z = \cos kt$ y $z = \operatorname{sen} kt$, aparecen en el estudio de las vibraciones. Ésta es la razón por la que las curvas $u = \operatorname{sen} kt$

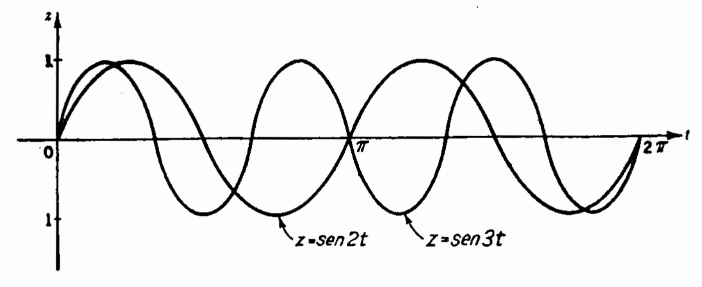


FIG. 280.

y $u = \cos kt$ (Fig. 280) constituyen el fundamento de los mecanismos vibratorios. Debe hacerse constar que la ecuación diferencial [10] representa el caso ideal de no existir rozamiento o resistencia. Esta última se hace intervenir en la ecuación diferencial del movimiento vibratorio añadiendo otro término rz',

$$z'' + rz' + k^2z = 0, \qquad [11]$$

cuyas soluciones son vibraciones «amortiguadas», que se expresan matemáticamente mediante las fórmulas

$$e^{-rt/2} \cos \omega t, \; e^{-rt/2} \operatorname{sen} \omega t; \qquad \omega = \sqrt{k^2 - \left(\frac{r}{2}\right)^2},$$

y se hallan representadas gráficamente en la figura 281. (Como ejercicio, el lector verificará esas soluciones, llevando a cabo las respecti-

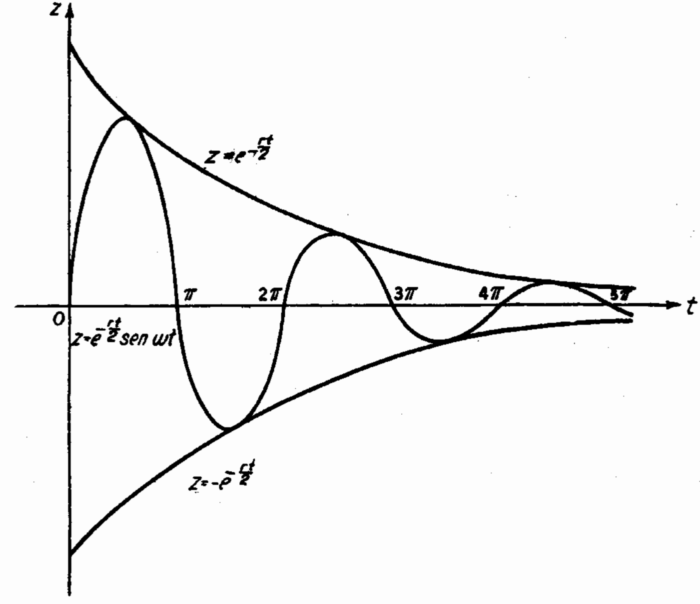


FIG. 281.—Vibraciones amortiguadas.

vas derivaciones.) En este caso, las oscilaciones son del mismo tipo que las del seno o coseno; pero su intensidad disminuye por efecto del factor exponencial, decreciendo más o menos rápidamente de acuerdo con la magnitud del coeficiente r, que representa el rozamiento.

4. Las leyes de la dinámica de Newton.—Aunque exceda a los límites señalados a este libro un análisis más detallado de estos hechos,

deseamos analizar el aspecto general de los conceptos fundamentales mediante los cuales revolucionó Newton la física y la mecánica. Si se considera el movimiento de una partícula de masa, m, y coordenadas espaciales en función del tiempo, t, $x(t)$, $y(t)$, $z(t)$, las componentes de la aceleración serán las derivadas segundas $x''(t)$, $y''(t)$, $z''(t)$, y el paso verdaderamente importante dado por Newton consistió en comprender que las cantidades mx'', my'', mz'', pueden considerarse como las componentes de la fuerza que actúa sobre la partícula. A primera vista, esto puede parecer sólo una nueva definición formal del concepto de «fuerza» en física. Pero la gran realización de Newton estriba en haber formulado su definición de acuerdo con los fenómenos reales de la Naturaleza, ya que ésta nos procura a menudo campos de fuerzas que nos son conocidos de antemano, sin saber nada acerca del movimiento particular que deseamos estudiar. El mayor triunfo de Newton en la dinámica, la justificación de las leyes de Kepler del movimiento de los planetas, pone claramente de manifiesto la armonía existente entre sus conceptos matemáticos y la Naturaleza. Newton empezó por suponer que la atracción de la gravedad es inversamente proporcional al cuadrado de la distancia; si admitimos que el Sol se encuentra en el origen del sistema de coordenadas y que un planeta dado tiene por coordenadas x, y, z, se deduce que las componentes de la fuerza en las direcciones x, y, z, son iguales, respectivamente, a

$$-k \cdot \frac{x}{r^3}, \qquad -k \cdot \frac{y}{r^3}, \qquad -k \cdot \frac{z}{r^3},$$

siendo k una constante gravitatoria independiente del tiempo, y $r = \sqrt{x^2 + y^2 + z^2}$, la distancia del Sol al planeta. Estas expresiones determinan un campo de fuerzas que no depende del movimiento de la partícula dentro del mismo. Este conocimiento del campo de fuerzas se combina ahora con la ley general de la dinámica de Newton (es decir, con la expresión de la fuerza en función del movimiento); igualando ambas expresiones, se tiene:

$$mx'' = \frac{-kx}{(x^2 + y^2 + z^2)^{3/2}},$$

$$my'' = \frac{-ky}{(x^2 + y^2 + z^2)^{3/2}},$$

$$mz'' = \frac{-kz}{(x^2 + y^2 + z^2)^{3/2}},$$

sistema de tres ecuaciones diferenciales con tres funciones desconocidas $x(t)$, $y(t)$, $z(t)$. Este sistema puede resolverse y resulta, de acuerdo

con las observaciones empíricas de Kepler, que la órbita del planeta es una cónica en uno de cuyos focos se encuentra el Sol, deduciéndose además que las áreas barridas por el radio vector que une el Sol con el planeta son iguales para tiempos iguales, y que los cuadrados de los períodos de una revolución completa de dos planetas son proporcionales a los cubos de sus distancias al Sol. Omitimos las correspondientes demostraciones.

El problema de las vibraciones nos procura un ejemplo más elemental del método de Newton. Supongamos que tenemos una partícula que se mueve a lo largo de una recta, el eje de las x, y que está unida al origen por una fuerza elástica tal como un resorte o una tira de caucho. Si se aparta la partícula de su posición de equilibrio en el origen hasta llevarla a una posición dada por su coordenada x, la fuerza la hará retroceder con una intensidad que supondremos proporcional a la elongación x; puesto que la fuerza está dirigida hacia el origen, vendrá representada por $-k^2x$, siendo $-k^2$ un factor negativo de proporcionalidad, que expresa la fuerza del resorte o de la tira de caucho. Supondremos además que existe rozamiento, el cual retarda el movimiento, y que éste es proporcional a la velocidad x' de la partícula, con otro factor de proporcionalidad $-r$. Entonces, la fuerza total en un instante cualquiera será $-k^2x - rx'$, y, de acuerdo con el principio general de Newton, encontramos que $mx'' = -k^2x - rx'$; o sea,

$$mx'' + rx' + k^2x = 0.$$

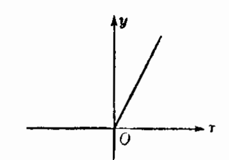
Ésta es precisamente la ecuación diferencial [11] de las vibraciones amortiguadas que ya hemos mencionado.

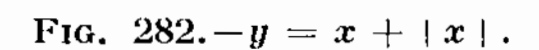
Este sencillo ejemplo es de gran importancia, puesto que numerosas clases de sistemas vibrantes, mecánicos y eléctricos, pueden describirse matemáticamente mediante dicha ecuación diferencial. He aquí, pues, un ejemplo típico, en el que una formulación matemática abstracta pone simultáneamente al descubierto la íntima estructura de muchos fenómenos, aparentemente distintos e inconexos. Esta abstracción, que, partiendo de la naturaleza particular de un fenómeno dado, llega hasta la formulación de la ley general que gobierna toda una clase de ellos, es uno de los rasgos característicos del tratamiento matemático de los fenómenos físicos.

SUPLEMENTO AL CAPÍTULO VIII

I. CUESTIONES DE PRINCIPIO

1. Derivabilidad.—Hemos establecido una conexión entre el concepto de derivada de una función $y = f(x)$ y la idea intuitiva de tangente a la gráfica de la función. Puesto que el concepto general de función es tan amplio, se hace necesario, en interés de la perfección lógica, prescindir de esa dependencia de la intuición geométrica, pues no tenemos ninguna seguridad de que los hechos intuitivos familiares, que resultan de la consideración de las curvas más sencillas, tales como circunferencias o elipses, subsistan necesariamente en las gráficas de funciones más complicadas. Consideremos, p. ej., la función de




FIG. 282.—$y = x + |x|$.

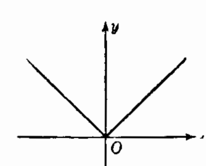


FIG. 283.—$y = |x|$.

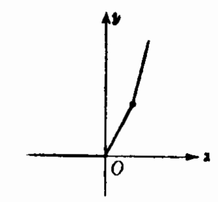


FIG. 284.—$y = x + |x| + (x - 1) + |x - 1|$.

la figura 282, cuya gráfica tiene un punto anguloso. Esta función viene definida por la ecuación $y = x + |x|$, en la cual $|x|$ es el valor absoluto de x; es decir,

$$y = x + x = 2x \quad \text{para} \quad x \geqslant 0,$$
$$y = x - x = 0 \quad \text{para} \quad x < 0.$$

Otro ejemplo lo proporciona la función $y = |x|$; otro más lo tenemos en la función siguiente: $y = x + |x| + (x - 1) + |x - 1|$. Las gráficas de estas funciones carecen en ciertos puntos de tangente definida, o de dirección, lo que significa que las funciones no poseen derivadas para los correspondientes valores de x.

Ejercicios:

1. Fórmese la función $f(x)$ cuya gráfica es la mitad de un hexágono regular.
2. ¿Dónde se encuentran los puntos angulosos de la gráfica de

$$f(x) = (x + |x|) + {}^1/_2 \{ (x - {}^1/_2) + |x - {}^1/_2| \} + {}^1/_4 \{ (x - {}^1/_4) + |x - {}^1/_4| \}?$$

¿Cuáles son las discontinuidades de $f'(x)$?

Otro ejemplo sencillo de no derivabilidad nos lo procura la función siguiente:

$$y = f(x) = x \operatorname{sen} \frac{1}{x},$$

que se obtiene de la función sen $1/x$ (véase pág. 294) multiplicando por el factor x. Definimos $f(x)$ como igual a cero para $x = 0$. Esta función, cuya gráfica para valores positivos de x aparece en la figura 285, es continua en todo punto. La gráfica oscila infinitas veces en el entorno de $x = 0$, haciéndose muy pequeñas las «ondas» al acercarse a ese punto. La pendiente de estas ondas está dada por

$$f'(x) = \operatorname{sen} \frac{1}{x} - \frac{1}{x} \cos \frac{1}{x}$$

(lo que el lector puede verificar como ejercicio); cuando x tiende a cero, esta pendiente oscila entre límites constantemente crecientes, positivos y negativos. Para $x = 0$, podemos intentar la determinación de la derivada como límite para $h \rightarrow 0$ del cociente de incrementos

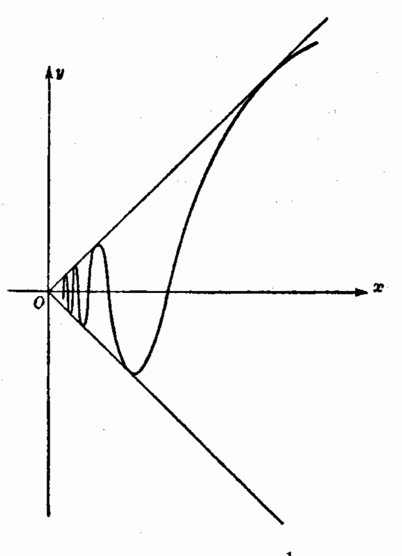


FIG. 285.—$y = x \operatorname{sen} \frac{1}{x}$

$$\frac{f(0 + h) - f(0)}{h} = \frac{h \operatorname{sen} \frac{1}{h}}{h} = \operatorname{sen} \frac{1}{h}$$

Pero al tender h a cero, este cociente oscila entre -1 y $+1$, sin tender a ningún límite, por lo que la función carece de derivada para $x=0$.

Estos ejemplos acusan la existencia de una dificultad intrínseca. Weierstrass ilustró de forma sorprendente esta situación construyendo una función continua cuya gráfica carece de tangente en todos sus puntos. En tanto que la derivabilidad implica la continuidad, estos ejemplos prueban que la continuidad no entraña la derivabilidad, puesto que la función de Weierstrass es continua y no tiene derivada en ninguno de sus puntos. En la práctica, esas dificultades no aparecen; excepto, tal vez, en puntos aislados, las curvas son tales que no

sólo es posible derivarlas, sino que además la derivada es continua. Entonces, ¿por qué no habríamos de prescribir la ausencia de estos fenómenos «patológicos» en los problemas que hemos de considerar? Esto es exactamente lo que se hace en el cálculo, donde sólo se consideran las funciones derivables. En el capítulo VIII hemos derivado muchos tipos de funciones, con lo que quedó demostrada su derivabilidad.

Puesto que la existencia de derivada de una función no es una consecuencia lógica del primer concepto, deberá demostrarse o suponerse. El concepto de tangente o de dirección de una curva, que constituyó originalmente la base del concepto de derivada, se deduce entonces de la pura definición analítica de esta última. Si la función $y = f(x)$ tiene derivada; es decir, si el cociente de incrementos $\frac{f(x+h) - f(x)}{h}$ tiene límite único $f'(x)$, cuando h tiende a cero por uno u otro lado, se dice que la curva correspondiente tiene una tangente con pendiente $f'(x)$. Queda así invertida la actitud ingenua de Fermat, Newton y Leibniz, en interés de la compatibilidad lógica.

Ejercicios:

1. Demuéstrese que la función continua definida por x^2 sen $(1/x)$ tiene derivada en $x = 0$.

2. Demuéstrese que la función arc tg $(1/x)$ es discontinua para $x = 0$; que x arc tg $(1/x)$ es continua, pero carece de derivada en $x = 0$, y que x^2 arc tg $(1/x)$ tiene derivada para $x = 0$.

2. La integral.—La situación es completamente análoga respecto a la integral de una función continua $f(x)$. En lugar de considerar el «área situada debajo de la curva» $y = f(x)$ como una cantidad que existe con toda evidencia y que puede expresarse *a posteriori* como límite de una suma, *definimos* la integral mediante este límite y consideramos el concepto de integral como la base fundamental a partir de la cual se deduce después el concepto general de área. Nos vemos obligados a proceder así debido a la vaguedad de la intuición geométrica cuando se aplica a conceptos analíticos tan generales como el de función continua. Comencemos formando la suma:

$$S_n = \sum_{j=1}^{n} f(v_j)(x_j - x_{j-1}) = \sum_{j=1}^{n} f(v_j)\,\Delta x_j, \qquad [1]$$

siendo $x_0 = a, x_1, \ldots, x_n = b$, una subdivisión del intervalo de integración; $\Delta x_j = x_j - x_{j-1}$ es el incremento de x, o longitud del intervalo de lugar j, y v_j un valor arbitrario de x en este subintervalo; es decir, $x_{j-1} \leqslant v_j \leqslant x_j$ (p. ej., podemos tomar $v_j = x_j$ o $v_j = x_{j-1}$).

Formemos ahora una sucesión de tales sumas, de forma que el número n de subintervalos aumente y al mismo tiempo tienda a cero la máxima amplitud de los mismos. El hecho fundamental es éste: la suma S_n, para una función continua dada, $f(x)$, tiende a un límite definido, A, que es independiente de la particular elección de los subintervalos y de los puntos v_j. Por definición, este límite es la integral $A = \int_a^b f(x)dx$.

Naturalmente, la existencia de este límite requiere una demostración analítica, si deseamos independizarnos de la noción geométrica intuitiva de área. Esta demostración puede consultarse en cualquier tratado riguroso de cálculo integral.

Al comparar los procesos de derivación e integración nos enfrentamos con la siguiente y contradictoria situación: la derivabilidad es, sin ninguna duda, una condición restrictiva respecto al concepto de función continua, pero la ejecución del proceso de derivación, esto es, el algoritmo del cálculo diferencial, es en la práctica un procedimiento directo basado en unas cuantas reglas sencillas. Por el contrario, toda función continua, sin excepción alguna, es integrable entre dos límites dados cualesquiera, y, en cambio, la formulación explícita de tales integrales, aun para las funciones más sencillas, es en general una tarea muy ardua. Por ello, en muchos casos, el teorema fundamental del cálculo integral constituye el instrumento decisivo para llevar a cabo la integración; sin embargo, para la mayoría de las funciones, aun para las más elementales, la integración no conduce a expresiones explícitas sencillas, y el cálculo numérico de las integrales requiere el uso de métodos menos elementales.

3. Otras aplicaciones del concepto de integral. Trabajo. Rectificación.—Al disociar la noción analítica de integral de su interpretación geométrica original nos encontramos con otras muchas interpretaciones y aplicaciones igualmente importantes; p. ej., la integral puede interpretarse en mecánica como expresión del concepto de trabajo, bastando para nuestro objeto la exposición del caso siguiente, elegido entre los más elementales: supongamos una masa en movimiento a lo largo del eje x bajo el influjo de una fuerza dirigida a lo largo del mismo. La masa se supone concentrada en el punto de abscisa x y la fuerza viene dada como una función $f(x)$ de la posición, indicándose con el signo de $f(x)$ si está dirigida en la dirección positiva o negativa del eje. Si la fuerza es constante y desplaza la masa desde el punto a al b, el trabajo realizado es igual al producto $(b - a)f$ de la intensidad f de la fuerza por la distancia recorrida por la masa; pero si la

intensidad varía con x, nos vemos precisados a definir la cantidad de trabajo realizado por un proceso de límite (como hicimos para definir la velocidad). A este objeto, dividimos el intervalo de extremos a y b en subintervalos menores, mediante los puntos $x_0 = a, x_1, ..., x_n = b$; a continuación suponemos que en cada subintervalo la fuerza se mantiene constante e igual, p. ej., a $f(x_\nu)$, que es su verdadero valor en el extremo del mismo, y después calculamos el trabajo que correspondería a esta fuerza que varía por saltos:

$$S_n = \sum_{\nu=1}^{n} f(x_\nu)\,\Delta x_\nu.$$

Si, como antes, hacemos ahora más fina la subdivisión, de manera que n tienda a infinito, vemos que la suma anterior tiende a la integral

$$\int_a^b f(x)\,dx.$$

De esta manera el trabajo realizado por una fuerza continuamente variable se define por medio de una integral.

Como nuevo ejemplo vamos a considerar una masa m sujeta por un muelle elástico al origen $x = 0$; de acuerdo con lo dicho en la página 471, la fuerza $f(x)$ será proporcional a x:

$$f(x) = -k^2 x,$$

donde k^2 es una constante positiva. El trabajo realizado por dicha fuerza cuando la masa se desplaza desde el origen a la posición $x = b$ será:

$$\int_0^b -k^2 x\,dx = -k^2 \frac{b^2}{2},$$

y el trabajo que debemos hacer contra tal fuerza para llevar el extremo del resorte a esta posición es $+\,k^2 \dfrac{b^2}{2}$.

Una segunda aplicación del concepto general de integral es la noción de longitud de un arco de curva. Supongamos que la porción de la curva que se considera está representada por una función $y = f(x)$, cuya derivada $f'(x) = \dfrac{dy}{dx}$ es también una función continua. Para definir su longitud procederemos en igual forma que si tuviéramos que medir la curva para algún fin práctico utilizando una vara de medir rígida. Inscribimos en el arco AB una poligonal de n lados convenientemente pequeños, y consideramos la longitud total L_n de

esta poligonal como una primera aproximación; si hacemos aumentar n al propio tiempo que tiende a cero la máxima longitud de los lados de la poligonal, definimos

$$L = \lim L_n$$

como la longitud del arco AB. (En el capítulo VI obtuvimos la longitud de una circunferencia de esta misma forma, como límite de los perímetros de polígonos regulares inscritos.) Puede demostrarse que para curvas no muy arbitrarias este límite existe y es independiente de la particular manera de elegir la sucesión de poligonales inscritas; estas curvas se dice que son *rectificables*. Cualquier curva «razonable» de las que aparecen en la teoría o en las aplicaciones será siempre rectificable y no hemos de preocuparnos de investigar los casos patológicos. Nos contentaremos con demostrar que todo arco AB, correspondiente a una función $y=f(x)$ con derivada continua $f'(x)$, es rectificable en el sentido dicho, y que su longitud L puede expresarse por una integral.

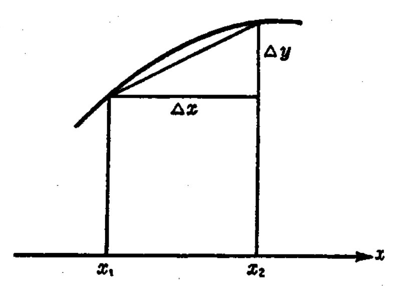


FIG. 286.—Longitud de un arco de curva.

A este fin, representemos las abscisas de A y B por a y b, respectivamente; subdividamos el intervalo (a, b) mediante los puntos $x = a, x_1, \ldots, x_j, \ldots, x_n = b$, y consideremos la poligonal de vértices $x_j, y_j = f(x_j)$ situados directamente por encima de estos puntos de subdivisión. La longitud de un lado cualquiera de la poligonal será $\sqrt{(x_j - x_{j-1})^2 + (y_j - y_{j-1})^2} = \sqrt{\Delta x_j^2 + \Delta y_j^2} = \Delta x_j \sqrt{1 + \left(\frac{\Delta y_j}{\Delta x_j}\right)^2}$. En consecuencia, tendremos como longitud total de la poligonal

$$L_n = \sum_{j=1}^{n} \sqrt{1 + \left(\frac{\Delta y_j}{\Delta x_j}\right)^2}\, \Delta x_j.$$

Si hacemos ahora tender n a infinito, los cocientes de incrementos $\frac{\Delta y_j}{\Delta x_j}$ tenderán a la derivada $\frac{dy}{dx} = f'(x)$ y obtenemos, como expresión de la longitud L, la integral

$$L = \int_a^b \sqrt{1 + [f'(x)]^2}\, dx. \qquad [2]$$

Sin entrar en más detalles de esta discusión teórica, vamos a terminar con dos notas suplementarias. Primera, si se considera que B es un punto variable de abscisa x sobre la curva, entonces $L = L(x)$ se transforma en una función de x, y tenemos por el teorema fundamental

$$L'(x) = \frac{dL}{dx} = \sqrt{1 + [f'(x)]^2},$$

fórmula de uso muy frecuente. Segunda, en tanto que la fórmula [2] da la solución «general» del problema, sólo procura una expresión explícita de la longitud del arco en casos particulares. Para esto tendremos que sustituir la función concreta $f(x)$, o mejor dicho, $f'(x)$, en [2] y a continuación realizar la integración efectiva de la expresión así obtenida, lo que entraña en general dificultades insuperables si nos restringimos a no hacer uso de otras funciones que las elementales, únicas consideradas en esta obra. Vamos a mencionar algunos casos en los que la integración es posible. La función

$$y = f(x) = \sqrt{1 - x^2}$$

representa la circunferencia unidad, y tenemos $f'(x) = \frac{dy}{dx} = -\frac{x}{\sqrt{1 - x^2}}$; o sea, $\sqrt{1 + f'(x)^2} = \frac{1}{\sqrt{1 - x^2}}$, de modo que la longitud de un arco de circunferencia está expresada por la integral

$$\int_a^b \frac{dx}{\sqrt{1 - x^2}} = \text{arc sen } b - \text{arc sen } a.$$

Para la parábola $y = x^2$, se tiene $f'(x) = 2x$, de donde la longitud del arco entre $x = 0$ y $x = b$ será:

$$\int_0^b \sqrt{1 + 4x^2}\, dx.$$

Para la curva $y = \log \text{sen } x$, tenemos $f'(x) = \cot x$, y la longitud del arco viene expresada por

$$\int_a^b \sqrt{1 + \cot^2 x}\, dx.$$

Nos contentaremos con escribir estas expresiones integrales, las cuales podrían ser calculadas disponiendo de algún recurso más de los que tenemos a mano, pero no deseamos ir más lejos por este camino.

II. ÓRDENES DE INFINITUD

1. La función exponencial y las potencias de «x».—Encontramos con frecuencia en matemáticas sucesiones a_n que tienden a infinito, y a menudo se precisa comparar una de tales sucesiones con otra b_n, también divergente, pero quizá «más rápidamente» que lo es a_n. Para precisar este concepto, diremos que b_n tiende a infinito con mayor rapidez que a_n, o que es de *mayor orden de infinitud* que a_n, si la razón a_n/b_n (en la cual tanto el numerador como el denominador tienden a infinito) tiende a cero al crecer n. Así, p. ej., la sucesión $b_n=n^2$ tiende a infinito más rápidamente que la sucesión $a_n=n$, y esta última a su vez lo hace con mayor rapidez que la $c_n = \sqrt{n}$; en efecto:

$$\frac{a_n}{b_n} = \frac{n}{n^2} = \frac{1}{n} \to 0; \qquad \frac{c_n}{a_n} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}} \to 0.$$

Es evidente que n^s tiende a infinito más rápidamente que n^r siempre que $s > r > 0$, ya que $n^r/n^s = 1/n^{(s-r)} \to 0$.

Si el cociente a_n/b_n tiende a un límite constante c, distinto de cero, diremos que las dos sucesiones a_n y b_n tienden a infinito con la misma rapidez, o que tienen el *mismo orden de infinitud*. Así, $a_n = n^2$ y $b_n = 2n^2 + n$ tienen el mismo orden de infinitud, dado que

$$\frac{a_n}{b_n} = \frac{n^2}{2n^2 + n} = \frac{1}{2 + \frac{1}{n}} \to \frac{1}{2}$$

Cabría pensar que utilizando las distintas potencias de n como patrones podrían medirse los distintos órdenes de infinitud de las diferentes sucesiones divergentes, a_n. Para ello habría que determinar una potencia conveniente n^s del mismo orden de infinitud que a_n; es decir, tal que a_n/n^s tendiera a una constante finita distinta de cero. Resulta notable el hecho de que esto no sea posible en todos los casos, pues *la función exponencial* a^n *con* $a > 1$ (p. ej., e^n) *tiende a infinito más de prisa que cualquier potencia* n^s, *por muy grande que se elija* s, *en tanto que* $\log n$ *tiende a infinito más despacio que cualquier potencia* n^s, *por pequeño que sea el exponente positivo* s. En otras palabras, se tienen las relaciones

$$\frac{n^s}{a^n} \to 0 \qquad [1]$$

y

$$\frac{\log n}{n^s} \to 0 \qquad [2]$$

para n tendiendo a infinito. Por supuesto que el exponente s no precisa ser entero, sino que puede ser cualquier número positivo fijo.

Para demostrar [1] simplificaremos en primer lugar el enunciado, considerando la raíz de índice s de dicho cociente, pues si esta raíz tiende a cero, lo mismo ocurrirá con el cociente dado. Por consiguiente, nos basta comprobar que

$$\frac{n}{a^{n/s}} \to 0$$

al tender n a infinito. Sea $b = a^{1/s}$; como se ha supuesto que a es mayor que 1, b y, por tanto, también $\sqrt{b} = b^{1/2}$ serán mayores que 1. Podemos escribir:

$$b^{1/2} = 1 + q,$$

siendo q positivo. Si tenemos ahora en cuenta la desigualdad [6] dada en la página 22, tendremos:

$$b^{n/2} = (1 + q)^n \geqslant 1 + nq > nq,$$

de modo que

$$a^{n/s} = b^n > n^2q^2$$

y

$$\frac{n}{a^{n/s}} < \frac{n}{n^2q^2} = \frac{1}{nq^2}$$

Con lo que la demostración queda efectuada, pues el segundo miembro de la última igualdad tiende a cero al tender n a infinito.

La relación

$$\frac{x^s}{a^x} \to 0 \qquad [3]$$

subsiste evidentemente si x tiende a infinito siguiendo los valores de una sucesión divergente cualquiera x_1, x_2, ..., que no precisa coincidir con la sucesión de los números naturales 1, 2, 3, ... En efecto, si $n - 1 \leqslant x \leqslant n$, se tiene:

$$\frac{x^s}{a^x} < \frac{n^s}{a^{n-1}} = a \cdot \frac{n^s}{a^n} \to 0.$$

Puede hacerse uso de esta observación para establecer [2]. Si escribimos $x = \log n$ y $e^s = a$, de forma que $n = e^x$ y $n^s = (e^s)^x$, el cociente que figura en [2] se transforma en

$$\frac{x}{a^x},$$

que es un caso especial de [3] para $s = 1$.

Ejercicios:

1. Demuéstrese que, para $x \to \infty$, la función log log x tiende a infinito más despacio que log x.

2. La derivada de $x/\log x$ es $1/\log x - 1/(\log x)^2$. Demuéstrese que, para valores grandes de x, es «asintóticamente» equivalente a su primer término, $1/\log x$; es decir, que su cociente tiende a 1 cuando $x \to \infty$.

2. Orden de infinitud de log (n!).—En muchas aplicaciones, p. ej., en la teoría de las probabilidades, es importante conocer el orden de infinitud o «comportamiento» asintótico de $n!$ para valores grandes de n. Nos conformaremos aquí con estudiar el logaritmo de $n!$; esto es, la expresión

$$P_n = \log 2 + \log 3 + \log 4 + \cdots + \log n.$$

Vamos a demostrar que el «valor asintótico» de P_n es igual a $n \log n$; es decir, que

$$\frac{\log(n!)}{n \log n} \to 1$$

cuando $n \to \infty$.

La demostración es un ejemplo típico de un método muy utilizado de comparar una suma con una integral. En la figura 287 la

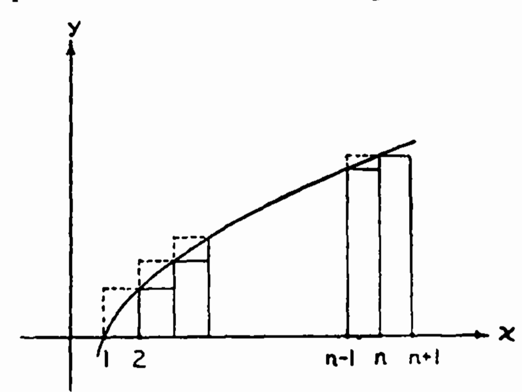


FIG. 287.—Cálculo aproximado de log ($n!$).

expresión P_n es igual a la suma de las áreas de los rectángulos trazados con línea llena, y que en su conjunto no exceden al área

$$\int_1^{n+1} \log x\, dx = (n+1)\log(n+1) - (n+1) + 1$$

comprendida por la curva logarítmica desde 1 a $n + 1$ (véase página 461, Ejercicio 1). Pero la suma P_n es también igual al área total

de los rectángulos que tienen parte de su contorno de trazos, y que en su conjunto superan al área situada bajo la curva entre 1 y n, la cual es igual a

$$\int_1^n \log x\,dx = n\log n - n + 1.$$

Tenemos así

$$n\log n - n + 1 < P_n < (n+1)\log(n+1) - n,$$

y dividiendo por $n \log n$,

$$1 - \frac{1}{\log n} + \frac{1}{n\log n} < \frac{P_n}{n\log n} < (1 + 1/n)\frac{\log(n+1)}{\log n} - \frac{1}{\log n} =$$

$$= (1 + 1/n)\frac{\log n + \log(1 + 1/n)}{\log n} - \frac{1}{\log n}$$

Evidentemente, ambas acotaciones tienden a 1 al tender n a infinito, con lo cual queda demostrado lo que deseábamos.

Ejercicio.—Demuéstrese que ambas acotaciones son mayores que $1 - 1/n$ y menores que $1 + 1/n$, respectivamente.

III. SERIES Y PRODUCTOS INFINITOS

1. Series funcionales.—Como ya hemos dicho, expresar un número s mediante una serie

$$s = b_1 + b_2 + b_3 + \cdots \qquad [1]$$

no es otra cosa que un simbolismo adecuado para expresar el hecho de que s es el límite, al tender n a infinito, de la sucesión de «sumas parciales»

$$s_1, s_2, s_3, \cdots,$$

donde

$$s_n = b_1 + b_2 + \cdots + b_n. \qquad [2]$$

En consecuencia, la ecuación [1] equivale a la igualdad límite

$$\lim s_n = s \quad \text{cuando} \quad n \to \infty, \qquad [3]$$

en la cual s_n está definida por [2]. Cuando existe el límite [3] decimos que la serie [1] *converge* hacia el valor s, mientras que si el límite [3] no existe, diremos que la serie *diverge*.

Por ejemplo, la serie

$$1 - {}^1/_3 + {}^1/_5 - {}^1/_7 + \cdots$$

converge hacia el valor $\pi/4$, y la serie

$$1 - {}^1/_2 + {}^1/_3 - {}^1/_4 + \cdots$$

converge hacia el valor log 2; por otra parte, la serie

$$1 - 1 + 1 - 1 + \cdots$$

es divergente (ya que sus sumas parciales toman alternativamente los valores 1 y 0), y la serie

$$1 + 1 + 1 + 1 + \cdots$$

diverge, por tender a infinito la sucesión de sus sumas parciales.

Nos hemos encontrado ya con series cuyos términos b_i son funciones de x de la forma

$$b_i = c_i x^i,$$

siendo constantes los coeficientes c_i. Dichas series se llaman *series potenciales*, y son límites de los polinomios que representan sus sumas parciales

$$S_n = c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n$$

(la adición del término constante c_0 sólo requiere un pequeño cambio en la notación [2]). Un desarrollo

$$f(x) = c_0 + c_1 x + c_2 x^2 + \cdots$$

de la función $f(x)$ en serie de potencias constituye así una forma de expresar aproximadamente $f(x)$ por medio de polinomios, que son las funciones más sencillas. Podemos resumir, completándolos, los resultados anteriores, escribiendo la sucesión siguiente de desarrollos en serie de potencias:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \cdots, \qquad \text{válido para } -1 < x < +1 \qquad [4]$$

$$\text{arc tg } x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots, \qquad \text{válido para } -1 \leq x \leq +1 \qquad [5]$$

$$\log(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots, \qquad \text{válido para } -1 < x \leq +1 \qquad [6]$$

$${}^1/_2 \log \frac{1+x}{1-x} = x + \frac{x^3}{3} + \frac{x^5}{5} + \cdots, \qquad \text{válido para } -1 < x < +1 \qquad [7]$$

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots, \qquad \text{válido para todo } x. \qquad [8]$$

A la relación anterior podemos añadir aún los importantes desarrollos siguientes:

$$\operatorname{sen} x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots, \quad \text{válido para todo } x; \qquad [9]$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots, \quad \text{válido para todo } x. \qquad [10]$$

La demostración es una consecuencia inmediata de las fórmulas (véase pág. 450)

$$\int_0^x \operatorname{sen} u \, du = 1 - \cos x, \qquad [a]$$

$$\int_0^x \cos u \, du = \operatorname{sen} x. \qquad [b]$$

Si partimos de la desigualdad $\cos x \leqslant 1$ y la integramos entre 0 y x, siendo x un número positivo dado, obtenemos (véase fórmula [13], pág. 423):

$$\operatorname{sen} x \leqslant x;$$

integrando de nuevo,

$$1 - \cos x \leqslant \frac{x^2}{2},$$

o lo que es equivalente,

$$\cos x \geqslant 1 - \frac{x^2}{2}$$

Si volvemos a integrar, obtenemos:

$$\operatorname{sen} x \geqslant x - \frac{x^3}{2 \cdot 3} = x - \frac{x^3}{3!}$$

Procediendo indefinidamente en igual forma, resultan las dos columnas de desigualdades

$$\operatorname{sen} x \leqslant x \qquad \cos x \leqslant 1$$

$$\operatorname{sen} x \geqslant x - \frac{x^3}{3!} \qquad \cos x \geqslant 1 - \frac{x^2}{2!}$$

$$\operatorname{sen} x \leqslant x - \frac{x^3}{3!} + \frac{x^5}{5!} \qquad \cos x \leqslant 1 - \frac{x^2}{2!} + \frac{x^4}{4!}$$

$$\operatorname{sen} x \geqslant x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} \qquad \cos x \geqslant 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!}$$

$$\cdots\cdots \qquad \cdots\cdots$$

Ahora bien: $x^n/n! \to 0$ al tender n a infinito; para probarlo, elijamos

un entero m tal que $x/m < {}^1/_2$, y escribamos $c = x^m/m!$; para cualquier entero $n > m$ (p. ej., $n = m + r$) se tiene:

$$0 < \frac{x^n}{n!} = c \cdot \frac{x}{m+1} \cdot \frac{x}{m+2} \cdots \frac{x}{m+r} < c({}^1/_2)^r,$$

y cuando $n \to \infty$, también $r \to \infty$ y, por tanto, $c\,({}^1/_2)^r \to 0$. En resumen,

$$\begin{cases} \operatorname{sen} x = x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \cdots \\ \cos x = 1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \dfrac{x^6}{6!} + \cdots \end{cases}$$

Por ser los términos de la serie alternativamente positivos y negativos y decrecientes (al menos para $|x| \leqslant 1$), resulta que el *error cometido al interrumpir la serie en un término cualquiera es inferior al valor absoluto del primer término despreciado.*

OBSERVACIONES: Estas series pueden utilizarse para el cálculo de tablas de líneas trigonométricas. Ejemplo: ¿Qué valor tiene sen 1º? La medida en radianes de 1º es $\pi/180$, y, en consecuencia,

$$\operatorname{sen}\frac{\pi}{180} = \frac{\pi}{180} - \frac{1}{6}\left(\frac{\pi}{180}\right)^3 + \cdots$$

El error cometido al prescindir de los términos siguientes es inferior a $\dfrac{1}{120}\left(\dfrac{\pi}{180}\right)^5$, que a su vez es menor que 0,00000000002. Por tanto, sen 1º = 0,0174524064, con diez cifras decimales exactas.

Para terminar, mencionaremos sin demostración la «serie binómica»

$$(1 + x)^a = 1 + ax + C_2^a x^2 + C_3^a x^3 + \cdots, \qquad [11]$$

en la cual C_s^a es el «coeficiente binómico»

$$C_s^a = \frac{a(a-1)(a-2)\cdots(a-s+1)}{s!}$$

Si $a = n$ es un entero positivo, $C_n^a = 1$, y para $s > n$ todos los coeficientes C_s^a de [11] son nulos. En este caso, la serie se reduce a la conocida fórmula del binomio. Fué una de las grandes aportaciones de Newton, realizada al comienzo de su carrera, el darse cuenta de que la fórmula elemental del binomio podía generalizarse para exponentes arbitrarios, positivos o negativos, racionales o irracionales. Cuan-

do a no es entero, el segundo miembro de [11] da lugar a una serie, válida para $-1 < x < +1$. Para $|x| > 1$, la serie [11] es divergente y la igualdad pierde su significado.

En particular, si sustituímos a por $^1/_2$ en [11], obtenemos el desarrollo

$$\sqrt{1+x} = 1 + \tfrac{1}{2}x - \frac{1}{2^2 \cdot 2!}x^2 + \frac{1 \cdot 3}{2^3 \cdot 3!}x^2 - \frac{1 \cdot 3 \cdot 5}{2^4 \cdot 4!}x^4 + \cdots \qquad [12]$$

Al igual que otros matemáticos del siglo XVIII, Newton no dió una demostración satisfactoria de la validez de su fórmula. El análisis adecuado de la convergencia y del intervalo de validez de este tipo de series no se hizo hasta el siglo XIX.

Ejercicio: Escríbanse las series de potencias de $\sqrt{1-x^2}$ y $1/\sqrt{1-x}$.

Los desarrollos [4] a [11] son casos especiales de la fórmula general de Brook Taylor (1685-1731), por la cual se consigue desarrollar una amplia clase de funciones $f(x)$ en forma de serie potencial

$$f(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + \cdots, \qquad [13]$$

determinando una ley que expresa los coeficientes c_i por medio de la función f y sus sucesivas derivadas. No es posible en este lugar detenernos a dar una demostración rigurosa de la fórmula de Taylor, formulando y estableciendo las condiciones requeridas para su validez. Nos limitaremos a algunas consideraciones que aclaran ciertas relaciones existentes entre importantes hechos matemáticos.

Para comenzar, supongamos posible el desarrollo [13], y admitamos además que $f(x)$ puede ser reiteradamente derivada, asegurándonos la existencia de la sucesión indefinida de derivadas

$$f'(x),\ f''(x),\ \cdots,\ f^{(n)}(x),\ \cdots$$

Finalmente, admitiremos también que una serie de potencias puede derivarse término a término en la misma forma que un polinomio. Con estas hipótesis, podemos determinar los coeficientes c_n a partir del conocimiento de la forma de comportarse $f(x)$ en el entorno de $x = 0$. Sustituyendo $x = 0$ en [13] encontramos

$$c_0 = f(0),$$

por anularse todos los términos de la serie que contienen x. Si ahora derivamos [13], resulta:

$$f'(x) = c_1 + 2c_2x + 3c_3x^2 + \cdots + nc_nx^{n-1} + \cdots \qquad [13']$$

Sustituyendo nuevamente $x = 0$, pero esta vez en [13′] y no en [13], obtenemos:

$$c_1 = f'(0).$$

Por derivación de [13′] resulta:

$$f''(x) = 2c_2 + 2 \cdot 3 \cdot c_3 x + \cdots + (n-1) \cdot n \cdot c_n x^{n-2} + \cdots ; \qquad [13'']$$

la cual, después de hacer $x = 0$, nos da

$$2!\, c_2 = f''(0).$$

De forma análoga, si se deriva [13″] y se hace $x = 0$,

$$3!\, c_3 = f'''(0),$$

y continuando este proceso, llegamos a la fórmula general

$$c_n = \frac{1}{n!} f^{(n)}(0),$$

en la cual $f^{(n)}(0)$ es el valor de la derivada n-ésima de $f(x)$ para $x=0$. Obtenemos así la *serie de Taylor*

$$f(x) = f(0) + xf'(0) + \frac{x^2}{2!} f''(0) + \frac{x^3}{3!} f'''(0) + \cdots \qquad [14]$$

Como ejercicio de derivación puede comprobar el lector que en los ejemplos [4] a [11] se satisface la ley de formación de los coeficientes que corresponde a la fórmula de Taylor.

2. Fórmula de Euler: $\cos x + i \operatorname{sen} x = e^{ix}$.—Uno de los resultados más fascinantes de las manipulaciones formalistas de Euler es la relación íntima existente en el campo complejo entre las funciones seno y coseno, de un lado, y la exponencial, por otro. Debe advertirse de antemano que la «demostración» de Euler y nuestra subsiguiente argumentación carecen del rigor necesario, y sólo constituyen ejemplos típicos de las manipulaciones formales del siglo XVIII.

Comencemos con la fórmula de De Moivre establecida en el capítulo II:

$$(\cos n\varphi + i \operatorname{sen} n\varphi) = (\cos \varphi + i \operatorname{sen} \varphi)^n.$$

Si en ésta hacemos la sustitución $\varphi = x/n$, obtenemos la fórmula

$$(\cos x + i \operatorname{sen} x) = \left(\cos\frac{x}{n} + i \operatorname{sen}\frac{x}{n}\right)^n$$

Si suponemos dado x, $\cos\frac{x}{n}$ diferirá muy poco de $\cos 0 = 1$ para valores grandes de n. Además, ya que

$$\frac{\operatorname{sen}\frac{x}{n}}{\frac{x}{n}} \to 1 \quad \text{cuando} \quad \frac{x}{n} \to 0$$

(véase pág. 318), vemos que $\operatorname{sen}\frac{x}{n}$ es asintóticamente igual a $\frac{x}{n}$, por lo que resulta justificado escribir la fórmula límite

$$\cos x + i \operatorname{sen} x = \lim\left(1 + \frac{ix}{n}\right)^n \text{ cuando } \quad n \to \infty . \qquad [14]$$

Al comparar el segundo miembro de esta igualdad con la relación (pág. 459)

$$e^z = \lim\left(1 + \frac{z}{n}\right)^n \text{ cuando } n \to \infty ,$$

obtenemos:

$$\cos x + i \operatorname{sen} x = e^{ix}, \qquad [15]$$

que constituye la fórmula de Euler.

Se puede llegar al mismo resultado por otro camino formal partiendo del desarrollo de e^z:

$$e^z = 1 + \frac{z}{1!} + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots ,$$

y haciendo en él $z = ix$, donde x es un número real. Si recordamos que las sucesivas potencias de i son i, -1, $-i$, 1, y así periódicamente, separando las partes real e imaginaria encontramos:

$$e^{ix} = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots\right) + i\left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots\right),$$

que, comparada con los segundos miembros de las series de $\operatorname{sen} x$ y $\cos x$, nos da nuevamente la fórmula de Euler.

Esta forma de razonar no constituye en absoluto una demostración de la relación [15]. La objeción que cabe hacer a nuestra segunda argumentación es que el desarrollo en serie de e^z fué deducido en la hipótesis de que z era real, por lo que requiere una justificación la sustitución $z = ix$. De forma análoga, la validez de la primera argumentación queda destruída por el hecho de que la fórmula

$$e^z = \lim (1 + z/n)^n \quad \text{cuando} \quad n \to \infty$$

fué obtenida en la hipótesis de ser z real.

Para liberar a la fórmula de Euler de la esfera del mero formalismo y darle una justificación rigurosa, conforme requiere toda verdad matemática, se precisa de los resultados de la teoría de funciones de variable compleja, una de las mayores realizaciones matemáticas del siglo XIX, y cuyo desarrollo fué también estimulado por otros muchos problemas. Hemos visto, p. ej., que los desarrollos de las funciones en serie de potencias convergen para intervalos de x diferentes, y cabe preguntarse por qué algunos de estos desarrollos convergen siempre, es decir, para todo x, en tanto que otros carecen de significado para $|x| > 1$.

Consideremos como ejemplo la serie geométrica [4] (pág. 483) que converge para $|x| < 1$. El primer miembro de esta igualdad tiene perfecto significado para $x = 1$, tomando el valor $\frac{1}{1+1} = \frac{1}{2}$, en tanto que la serie que figura en el segundo miembro se comporta de forma bastante más extraña, transformándose en

$$1 - 1 + 1 - 1 + \cdots$$

Esta serie no converge, ya que sus sumas parciales oscilan entre 1 y 0; todo lo cual indica la existencia de funciones que pueden dar lugar a series (divergentes) aunque por sí mismas no presenten ninguna irregularidad. Por supuesto que la función $\frac{1}{1+x}$ tiende a infinito cuando $x \to -1$. Según puede demostrarse fácilmente, la convergencia de una serie potencial para $x = a > 0$ supone siempre la convergencia para $-a < x < a$, de lo cual cabe deducir una «explicación» de la extraña conducta del desarrollo en $x = -1$, punto de discontinuidad de $\frac{1}{1+x}$. En cambio, la función $\frac{1}{1+x^2}$ se desarrolla en la serie

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \cdots$$

sin más que sustituir x por x^2 en [4]. Esta serie converge también para $|x| < 1$, mientras que para $x = 1$ da otra vez lugar a la serie divergente $1 - 1 + 1 - 1 + 1 \ldots$ y para $|x| > 1$ diverge explosivamente, aunque la propia función es regular en todo punto.

Resulta que una explicación completa de tales fenómenos únicamente es posible cuando las funciones se estudian no sólo para valores reales, sino también *complejos*, de la variable x; p. ej., la serie

correspondiente a $\frac{1}{1+x^2}$ debe divergir para $x = i$ a causa de que el denominador de la fracción se anula para dicho valor, y resulta que también debe divergir para todo x tal que $|x| > |i| = 1$, porque, según puede demostrarse, la convergencia para uno de tales valores de x supondría la convergencia para $x = i$. Así el problema de la convergencia de las series, que fué completamente olvidado en la primera época del desarrollo del cálculo, se transformó en uno de los factores principales que contribuyeron a la creación de la teoría de funciones de variable compleja.

3. La serie armónica y la función ζ. Producto de Euler.—Resultan particularmente interesantes algunas series cuyos términos son combinaciones sencillas de los números enteros; entre éstas vamos a considerar la «serie armónica»

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{n} + \cdots, \qquad [16]$$

que difiere de la correspondiente a log 2 en los signos de los términos pares.

La convergencia de la serie depende de si la sucesión

$$s_1, s_2, s_3, \cdots,$$

en la cual

$$s_n = 1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}, \qquad [17]$$

tiende a un límite finito. Aunque los términos de la serie [16] tienden a 0, es fácil ver que la serie no converge; en efecto, vamos a demostrar que si tomamos un número suficiente de términos su suma llega a exceder a cualquier número positivo, de forma que s_n crece sin limitación y, por tanto, la serie [16] «diverge hacia infinito». Para ver esto observamos que

$$s_2 = 1 + {}^1/_2,$$
$$s_4 = s_2 + ({}^1/_3 + {}^1/_4) > s_2 + ({}^1/_4 + {}^1/_4) = 1 + {}^2/_2,$$
$$s_8 = s_4 + ({}^1/_5 + \cdots + {}^1/_8) > s_4 + ({}^1/_8 + \cdots + {}^1/_8) = s_4 + {}^1/_2 > 1 + {}^3/_2,$$

y, en general,

$$s_{2^m} > 1 + \frac{m}{2} \qquad [18]$$

Así, p. ej., la suma parcial s_{2^m} es mayor que 100 en cuanto $m \geqslant 200$.

Aunque la serie armónica no converge, la serie

$$1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots + \frac{1}{n^s} + \cdots \qquad [19]$$

se demuestra que es convergente para todo valor de s mayor que 1, y define, para todo $s > 1$, la llamada función zeta:

$$(s) = \text{lím}\left(1 + \frac{1}{2^s} + \frac{1}{3^s} + \frac{1}{4^s} + \cdots + \frac{1}{n^s}\right) \quad \text{cuando} \quad n \to \infty, \qquad [20]$$

como función de la variable s. Existe una importante relación entre esta función y los números primos, la cual puede establecerse utilizando nuestros conocimientos acerca de la serie geométrica. Sea $p=2, 3, 5, 7, \ldots$, un número primo cualquiera; para $s \geqslant 1$ se tiene:

$$0 < \frac{1}{p^s} < 1,$$

de forma que

$$\frac{1}{1 - 1/p^s} = 1 + \frac{1}{p^s} + \frac{1}{p^{2s}} + \frac{1}{p^{3s}} + \cdots$$

Multipliquemos todas estas expresiones, correspondientes a los sucesivos números primos $p = 2, 3, 5, 7, \ldots$, sin preocuparnos de la validez de tal operación. En el primer miembro obtenemos el «producto» infinito

$$\left(\frac{1}{1 - 1/2^s}\right) \cdot \left(\frac{1}{1 - 1/3^s}\right) \cdot \left(\frac{1}{1 - 1/5^s}\right) \cdots =$$

$$= \text{límite para } n \to \infty \text{ de } \left[\frac{1}{1 - 1/p_1^s} \cdots \frac{1}{1 - 1/p_n^s}\right],$$

mientras que en el segundo aparece la serie

$$1 + \frac{1}{2^s} + \frac{1}{3^s} + \cdots = \zeta(s),$$

debido al hecho de que todo número entero mayor que 1 puede ser expresado de manera única como producto de potencias de números primos diferentes. Tenemos así representada la función zeta mediante el producto

$$\zeta(s) = \left(\frac{1}{1 - 1/2^s}\right)\left(\frac{1}{1 - 1/3^s}\right)\left(\frac{1}{1 - 1/5^s}\right) \cdots \qquad [21]$$

Si sólo existiera un número finito de primos distintos, $p_1, p_2, p_3, \ldots, p_r$, el producto que figura en el segundo miembro de [21] sería un producto finito ordinario y tendría, por tanto, un valor finito incluso para $s = 1$. Pero, según hemos visto, la serie zeta para $s = 1$,

$$\zeta(1) = 1 + 1/2 + 1/3 + \cdots,$$

es divergente. Este razonamiento, que puede transformarse sin dificultad en una demostración rigurosa, prueba la existencia de infinitos números primos. Por supuesto que esta demostración es más sofística e indirecta que la dada por Euclides (véase pág. 29). Ofrece la fascinación de un difícil ascenso a un pico montañoso al que puede llegarse desde el otro lado siguiendo una cómoda carretera.

Los productos infinitos tales como [21] son a veces tan útiles como las series para la representación de las funciones. Otro producto infinito, cuyo descubrimiento es una más de las aportaciones de Euler a la matemática, se refiere a la función trigonométrica sen x. Para comprender el desarrollo que sigue, comenzaremos con una observación relativa a las funciones enteras. Si $f(x) = a_0 + a_1x + ... + a_nx^n$ es un polinomio de grado n con n ceros distintos, $x_1, ..., x_n$, se sabe por álgebra que $f(x)$ admite la siguiente descomposición en factores lineales:

$$f(x) = a_n(x - x_1) \cdots (x - x_n)$$

(véase pág. 110). Sacando factor común el producto $x_1\, x_2 \ldots x_n$, podemos escribir:

$$f(x) = C\left(1 - \frac{x}{x_1}\right)\left(1 - \frac{x}{x_2}\right) \cdots \left(1 - \frac{x}{x_n}\right),$$

siendo C una constante cuyo valor, $C = a_0$, se obtiene sin más que hacer $x = 0$. Si en lugar de polinomios consideramos funciones $f(x)$ más complicadas, surge la cuestión de si es posible todavía una descomposición en producto de $f(x)$ mediante sus ceros. (Que esto en general no es posible se ve considerando el ejemplo de la función exponencial, la cual carece de ceros, pues $e^x \neq 0$ para todo x). Euler descubrió que para la función seno tal descomposición *es* posible. Con el fin de establecer la fórmula del modo más sencillo, consideremos sen πx en lugar de sen x. Esta función tiene los ceros $x = 0, \pm 1, \pm 2, \pm 3, ...$ ya que sen $\pi n = 0$ para todo valor entero n, y sólo para ellos. La fórmula de Euler dice que

$$\operatorname{sen} \pi x = \pi x \left(1 - \frac{x^2}{1^2}\right)\left(1 - \frac{x^2}{2^2}\right)\left(1 - \frac{x^2}{3^2}\right)\left(1 - \frac{x^2}{4^2}\right) \cdots \qquad [22]$$

Este producto infinito converge para todo x y es una de las fórmulas más bellas de la matemática. Para $x = {}^1/_2$ se obtiene:

$$\operatorname{sen}\frac{\pi}{2} = 1 = \frac{\pi}{2}\left(1 - \frac{1}{2^2 \cdot 1^2}\right)\left(1 - \frac{1}{2^2 \cdot 2^2}\right)\left(1 - \frac{1}{2^2 \cdot 3^2}\right)\left(1 - \frac{1}{2^2 \cdot 4^2}\right) \cdots$$

Si escribimos

$$1 - \frac{1}{2^2 \cdot n^2} = \frac{(2n-1)(2n+1)}{2n \cdot 2n},$$

obtenemos el producto de Wallis

$$\frac{\pi}{2} = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \cdots,$$

mencionado en la página 311.

Las demostraciones de todos estos resultados pueden consultarse en los textos de cálculo (véanse también págs. 519-520).

**IV. EL TEOREMA DE LOS NÚMEROS PRIMOS DEDUCIDO POR MÉTODOS ESTADÍSTICOS

Al aplicar los métodos matemáticos al estudio de los fenómenos naturales, nos quedamos satisfechos de ordinario con argumentaciones en cuyo desarrollo se interrumpe la concatenación del razonamiento lógico estricto mediante hipótesis más o menos plausibles. Incluso en la matemática pura se encuentran razonamientos que, si bien no procuran una demostración rigurosa, sugieren, no obstante, la solución correcta y señalan la dirección en que debe buscarse una demostración precisa. La solución de Bernoulli del problema de la braquistocrona (véase pág. 393) tiene este carácter, como así ocurre con la mayor parte de los primeros trabajos en el desarrollo del análisis matemático.

Por medio de un procedimiento típico de la matemática aplicada, y, en particular, de la mecánica estadística, vamos a presentar una argumentación que por lo menos hace plausible la verdad de la famosa ley de Gauss sobre la distribución de los números primos (un procedimiento análogo le fué sugerido a uno de los autores por el físico experimental Gustav Hertz). Este teorema, discutido empíricamente en el suplemento al capítulo I, dice que el número $A(n)$ de números primos no superiores a n es asintóticamente equivalente a $n/\log n$:

$$A(n) \sim \frac{n}{\log n}$$

Esto significa que el cociente de $A(n)$ por $n/\log n$ tiende al límite 1 cuando n tiende a infinito.

Comencemos haciendo la hipótesis de la *existencia* de una ley matemática que describa la distribución de los números primos en el sentido siguiente: para valores grandes de n, la función $A(n)$ es aproximadamente igual a la integral $\int_2^n W(x)\,dx$, siendo $W(x)$ una función

que mide la «densidad» de los números primos. (Elegimos 2 como límite inferior de la integral, por la razón de que para $x < 2$ evidentemente $A(x) = 0$.) Para precisar más, sean x y Δx dos números grandes, pero tales que el orden de magnitud de x sea superior al de Δx (p. ej., podemos convenir en que $\Delta x = \sqrt{x}$). Hecho esto, es posible decir que la distribución supuesta de los números primos es tal que el número de primos contenidos en el intervalo $[x, x + \Delta x]$ es aproximadamente igual a $W(x) \cdot \Delta x$ y, además, que $W(x)$ es una función de x cuya variación es tan lenta que la integral $\int_2^n W(x)\,dx$ puede sustituirse por una aproximación rectangular sin que varíe su valor asintótico. Después de estas observaciones preliminares, estamos en condiciones de comenzar el razonamiento.

Hemos demostrado ya (pág. 481) que para valores grandes, $\log n!$ es asintóticamente igual a $n \log n$:

$$\log n! \sim n \cdot \log n.$$

Vamos ahora a dar una segunda fórmula para $\log n!$ relacionada con los números primos y a comparar las dos expresiones. Veamos el número de veces que un número arbitrario p menor que n se halla contenido como factor en el entero $n! = 1 \cdot 2 \cdot 3 \cdots n$. Representaremos por $[a]_p$ el máximo entero k tal que p^k divide a a. Por ser única la descomposición en factores primos de cualquier entero, resulta que $[ab]_p = [a]_p + [b]_p$ para dos enteros cualesquiera a y b; por consiguiente,

$$[n!]_p = [1]_p + [2]_p + [3]_p + \cdots + [n]_p.$$

Los términos de la sucesión 1, 2, 3, ..., n, que son divisibles por p^k son p^k, $2p^k$, $3p^k$, ...; su número N_k para valores grandes de n es aproximadamente n/p^k. El número M_k de estos términos que son divisibles por p^k y no lo son por potencias mayores de p es igual a $N_k - N_{k+1}$; por tanto,

$$\begin{aligned}[n!]_p &= M_1 + 2M_2 + 3M_3 + \cdots \\ &= (N_1 - N_2) + 2(N_2 - N_3) + 3(N_3 - N_4) + \cdots \\ &= N_1 + N_2 + N_3 + \cdots \\ &= \frac{n}{p} + \frac{n}{p^2} + \frac{n}{p^3} + \cdots = \frac{n}{p-1}\end{aligned}$$

(Estas igualdades son, naturalmente, sólo aproximadas.)

Resulta que para valores grandes de n el valor de $n!$ viene dado,

aproximadamente, por el producto de todas las expresiones $p^{\frac{n}{p-1}}$ para todos los primos $p < n$. Tenemos así la fórmula

$$\log n! \sim \sum_{p<n} \frac{n}{p-1} \log p.$$

Al comparar esta con la relación asintótica anterior de $\log n!$, obtenemos, sustituyendo n por x,

$$\log x \sim \sum_{p<x} \frac{\log p}{p-1} \tag{1}$$

El paso siguiente y decisivo consiste en obtener una expresión asintótica en función de $W(x)$ para el segundo miembro de [1]. Como x es muy grande, podemos subdividir el intervalo comprendido entre 2 y $x = n$ en un gran número r de subintervalos mediante la elección de los puntos $2 = \xi_1, \xi_2, \ldots, \xi_r, \xi_{r+1} = x$, con los correspondientes incrementos $\Delta\xi_j = \xi_{j+1} - \xi_j$. En cada subintervalo pueden existir números primos, y todos los existentes en el j-ésimo subintervalo tendrán aproximadamente el valor ξ_j. Dada nuestra hipótesis acerca de $W(x)$, existen aproximadamente $W(\xi_j) \cdot \Delta\xi_j$ números primos en el subintervalo j-ésimo, y, en consecuencia, la suma del segundo miembro de [1] es aproximadamente igual a

$$\sum_{j=1}^{r+1} W(\xi_j) \frac{\log \xi_j}{\xi_j - 1} \cdot \Delta\xi_j.$$

Reemplazando esta suma finita por la integral que la aproxima, obtenemos como consecuencia plausible de [1] la relación

$$\log x \sim \int_2^x W(\xi) \frac{\log \xi}{\xi - 1} d\xi. \tag{2}$$

A partir de esta última vamos a determinar la función desconocida $W(x)$. Si sustituímos el signo $\sim$ por el ordinario de igualdad y derivamos ambos miembros respecto a x, por el teorema fundamental del cálculo se obtiene:

$$\frac{1}{x} = W(x) \frac{\log x}{x-1};$$
$$W(x) = \frac{x-1}{x \log x} \tag{3}$$

Hemos supuesto al comienzo de nuestra discusión que $A(x)$ es aproximadamente igual a $\int_2^x W(x)\,dx$; por tanto, $A(x)$ viene dado aproximadamente por la integral

$$\int_2^x \frac{x-1}{x \log x}\,dx. \qquad [4]$$

Para calcular esta integral observemos que la derivada de la función $f(x) = x/\log x$ es

$$f'(x) = \frac{1}{\log x} - \frac{1}{(\log x)^2}$$

Para valores grandes de x, las dos expresiones

$$\frac{1}{\log x} - \frac{1}{(\log x)^2}, \qquad \frac{1}{\log x} - \frac{1}{x \log x}$$

son aproximadamente iguales, pues para tales valores de x el sustraendo en ambas diferencias es mucho menor que el minuendo. En consecuencia, la integral [4] será asintóticamente igual a la integral

$$\int_2^x f'(x)\,dx = f(x) - f(2) = \frac{x}{\log x} - \frac{2}{\log 2},$$

ya que los integrandos son casi iguales a lo largo de la mayor parte del intervalo de integración. El término $2/\log 2$ puede despreciarse para valores grandes de x por tener valor constante, con lo que obtenemos finalmente el resultado

$$A(x) \sim \frac{x}{\log x},$$

que es la expresión del teorema citado sobre números primos.

No es posible pretender que la argumentación precedente sea más que una mera sugerencia, aunque un análisis más detallado pone de manifiesto el siguiente hecho: no es difícil dar una justificación completa de todos los pasos que con tanto atrevimiento hemos dado; en particular, para la ecuación [1], para la igualdad asintótica entre esta suma y la integral [2], y para el paso que conduce de [2] a [3]. Bastante más difícil es probar la *existencia* de la función de densidad $W(x)$ que hemos supuesto al principio; pero una vez hecha esta hipótesis, el *cálculo* de la función es cuestión relativamente sencilla. Desde este punto de vista, la demostración de la existencia de tal función representa la dificultad fundamental del problema de los números primos.

APÉNDICE

OBSERVACIONES SUPLEMENTARIAS PROBLEMAS Y EJERCICIOS

Muchos de los problemas que siguen a continuación están pensados para el lector con cierta formación matemática, y su intención no es tanto la de desarrollar una técnica rutinaria como la de estimular la habilidad inventiva.

Aritmética y Álgebra.

1. ¿Cómo veríamos que 3 no divide a ninguna potencia de 10, según se afirma en la página 70? (véase pág. 54).

2. Demuéstrese que el principio del mínimo entero es una consecuencia del teorema de inducción matemática (véase pág. 26).

3. Haciendo uso del teorema del binomio, aplicado al desarrollo de $(1 + 1)^n$, demuéstrese que $C_0^n + C_1^n + C_2^n + ... + C_n^n = 2^n$.

*4. Considérese un entero, $z = abc ...$, y fórmese la suma de sus cifras $a + b + c + ...$ Réstese ésta de z, táchese una de las cifras del resultado y sea w la suma de las restantes cifras. Conociendo solamente w, ¿puède darse una regla para determinar la cifra tachada? (Existe un caso ambiguo cuando $w = 0$.) Al igual que de otras muchas propiedades sencillas de las congruencias, puede hacerse uso de ésta como pasatiempo curioso.

5. Una progresión aritmética de primer orden es una sucesión de números, a, $a + d$, $a + 2d$, $a + 3d$, ..., tales que la diferencia entre dos términos consecutivos es constante. Una progresión aritmética de segundo orden es una sucesión de números a_1, a_2, a_3, ..., con la propiedad de que las diferencias $a_{i+1} - a_i$ forman una progresión aritmética de primer orden. Análogamente, una progresión aritmética de orden k es una sucesión tal que dichas diferencias forman una progresión aritmética de orden $k - 1$. Demuéstrese que los cuadrados de los números enteros constituyen una progresión aritmética de segundo orden, y establézcase por inducción que las potencias de exponente k de dichos números forman una progresión aritmética de orden k. Demuéstrese que cualquier sucesión cuyo n-ésimo término, a_n, viene dado por la expresión $c_0 + c_1 n + c_2 n^2 + ... + c_k n^k$, en la cual las c_i son constantes, es una progresión aritmética de orden k. Pruébese el teorema recíproco de este último para $k=2$, $k=3$, y, en general, para cualquier valor de k.

6. Demuéstrese que la suma de los n primeros términos de una progresión aritmética de orden k es a su vez una progresión aritmética de orden $k+1$.

7. ¿Cuántos divisores tiene el número 10296? (véase pág. 32).

8. Utilizando la identidad algebraica $(a^2+b^2)(c^2+d^2)=(ac-bd)^2+(ad+bc)^2$, demuéstrese por inducción que cualquier número entero $r = a_1 a_2 \dots a_n$, donde todas las a_i son sumas de dos cuadrados, es asimismo una suma de dos cuadrados. Compruébese esto con $2 = 1^2 + 1^2$, $5 = 1^2 + 2^2$, $8 = 2^2 + 2^2$, etc., para $r = 160$, $r = 1600$, $r = 1300$ y $r = 625$. Obténganse, si es posible, diferentes representaciones de estos números como sumas de dos cuadrados.

9. Aplíquese el resultado del ejercicio anterior para construir nuevas ternas de números pitagóricos partiendo de una terna dada.

10. Establézcanse reglas de divisibilidad, análogas a las dadas en la página 42, para los sistemas de numeración de bases 7, 11 y 12.

11. Pruébese que para dos números racionales positivos, $r = a/b$ y $s = c/d$, la desigualdad $r > s$ equivale a la $ac - bd > 0$.

12. Demuéstrese que para r y s positivos y $r < s$, se verifica siempre que

$$r < \frac{r+s}{2} < s \quad \text{y} \quad \frac{2}{[(1/r)+(1/s)]^2} < 2rs < (r+s)^2.$$

13. Siendo z un número complejo, pruébese por inducción que $z^n + 1/z^n$ puede expresarse como polinomio de grado n en la variable $w = z + 1/z$ (véase pág. 109).

*14. Utilizando la notación abreviada $\cos\varphi + i \operatorname{sen}\varphi = E(\varphi)$, se tiene $[E(\varphi)]^m = E(m\varphi)$. Utilícense esta fórmula y las dadas en la página 20 en relación con las progresiones geométricas (todas ellas conservan su validez para números complejos), para demostrar que

$$\operatorname{sen}\varphi + \operatorname{sen}2\varphi + \operatorname{sen}3\varphi + \cdots + \operatorname{sen}n\varphi = \frac{\cos\dfrac{\varphi}{2} - \cos(n+{}^1/_2)\varphi}{2\operatorname{sen}\dfrac{\varphi}{2}};$$

$${}^1/_2 + \cos\varphi + \cos2\varphi + \cos3\varphi + \cdots + \cos n\varphi = \frac{\operatorname{sen}(n+{}^1/_2)\varphi}{2\operatorname{sen}{}^1/_2\varphi}$$

15. Determínese la fórmula resultante de sustituir q por $E(\varphi)$ en la fórmula del ejercicio 3 de la página 25.

Geometría analítica.

Un estudio cuidadoso de los ejercicios que siguen, suplementado con dibujos y ejemplos numéricos, ayudará mucho a dominar los ele-

mentos de la geometría analítica. Se suponen conocidos las definiciones y los resultados más sencillos de la trigonometría.

Es con frecuencia útil imaginar una recta o un segmento como orientados de uno de sus puntos a otro. Por recta *orientada* PQ (o segmento *orientado* PQ) entendemos la recta (o segmento) que tiene el sentido de P a Q. En ausencia de una especificación explícita, una recta orientada l se supondrá con un sentido fijo arbitrario, salvo para el eje x, que se considerará siempre orientado de O a un punto de abscisa positiva, y análogamente para el eje orientado y. Las rectas orientadas (o los segmentos) se dirán paralelas únicamente cuando tienen el mismo sentido. El sentido de un segmento de una recta orientada se indicará atribuyendo signo más o menos a la distancia entre los extremos del segmento, según que éste tenga el mismo sentido que la recta o el opuesto. Es conveniente extender la terminología de «segmento PQ» al caso en que P y Q coinciden; a tal «segmento» se le asigna evidentemente longitud cero, pero carece de dirección.

16. Demuéstrese que si P_1 (x_1, y_1) y P_2 (x_2, y_2) son dos puntos cualesquiera, las coordenadas del punto medio, $P_0(x_0, y_0)$, del segmento P_1P_2 son: $x_0 = (x_1 + x_2)/2$, $y_0 = (y_1 + y_2)/2$. En general, demuéstrese que si P_1 y P_2 son distintos, las coordenadas del punto P_0 de la recta orientada P_1P_2 para el cual el cociente $P_1P_0 : P_1P_2$ de los segmentos orientados tiene el valor k, son:

$$x_0 = (1 - k)x_1 + kx_2, \qquad y_0 = (1 - k)y_1 + ky_2 .$$

(Varias rectas paralelas determinan segmentos proporcionales sobre dos transversales.)

Así, los puntos de la recta P_1P_2 tienen coordenadas de la forma $x = \lambda_1x_1 + \lambda_2x_2$, $y = \lambda_1y_1 + \lambda_2y_2$, con $\lambda_1 + \lambda_2 = 1$. Los valores $\lambda_1 = 1$ y $\lambda_1 = 0$ caracterizan los puntos P_1 y P_2, respectivamente. Los valores negativos de λ_2 corresponden a los puntos anteriores a P_1.

17. Caracterícese la posición de los puntos de una recta en forma análoga a la anterior, mediante los valores de k.

Es igualmente importante utilizar números positivos y negativos para indicar los sentidos de los giros en idéntica forma a como hemos hecho para las distancias. Por definición, el sentido de una rotación que lleve el eje orientado x a coincidir con el eje orientado y después de un giro de $90°$ se considera positivo. En el sistema usual de coordenadas, con el eje positivo x orientado hacia la derecha y el eje positivo y, hacia arriba, el sentido positivo corresponde al sentido de rotación contrario al de las agujas del reloj. Definimos el ángulo de dos rectas orientadas l_1 y l_2 como igual al ángulo que debe girar l_1

para quedar paralela a l_2. Naturalmente, este ángulo está determinado salvo un múltiplo entero de 360° (un giro completo). Así, el ángulo del eje orientado x con el eje orientado y es 90° o —270°, etc.

18. Si es α el ángulo que forma el eje orientado x con la recta orientada l; P_1 y P_2 dos puntos cualesquiera de l, y d la distancia orientada de P_1 a P_2, demuéstrese que

$$\cos\alpha = \frac{x_2 - x_1}{d}, \quad \operatorname{sen}\alpha = \frac{y_2 - y_1}{d}, \quad (x_2 - x_1)\operatorname{sen}\alpha = (y_2 - y_1)\cos\alpha.$$

Si la recta l no es perpendicular al eje x, la *pendiente* de l se define como

$$m = \operatorname{tg}\alpha = \frac{y_2 - y_1}{x_2 - x_1}$$

El valor de m no depende de la elección de sentido sobre la recta, ya que $\operatorname{tg}\alpha = \operatorname{tg}(\alpha + 180°)$, o lo que es equivalente: $(y_1 - y_2)/(x_1 - x_2) = (y_2 - y_1)/(x_2 - x_1)$.

19. Demuéstrese que la pendiente de una recta es nula, positiva o negativa según que una paralela a ella por el origen coincida con el eje x, esté en el primero y tercer cuadrantes, o en el segundo y cuarto, respectivamente.

En una recta orientada l distinguiremos el lado positivo del negativo de la forma siguiente: sea P un punto cualquiera exterior a l y Q el pie de la perpendicular a l por P. Diremos que P se encuentra en el lado positivo o negativo de l, según que el ángulo de l con la recta orientada QP sea 90° o —90°.

Vamos ahora a determinar la ecuación de una recta orientada l. Tracemos por el origen O una recta m perpendicular a l, y orientada de forma que el ángulo que forme con l sea de +90°. Al ángulo que forma el eje orientado x con m le llamaremos β; por tanto, $\alpha = 90° + \beta$, $\operatorname{sen}\alpha = \cos\beta$, $\cos\alpha = -\operatorname{sen}\beta$. Sea R, de coordenadas x_1 e y_1, el punto de intersección de m y l; representaremos por d la distancia orientada OR a la recta orientada m.

20. Demuéstrese que d es positivo si y sólo si O es un punto del lado negativo de l.

Tenemos: $x_1 = d\cos\beta$, $y_1 = d\operatorname{sen}\beta$ (compárese con ejercicio 18). Por tanto, $(x - x_1)\operatorname{sen}\alpha = (y - y_1)\cos\alpha$; o bien, $(x - d\cos\beta)\cos\beta = -(y - d\operatorname{sen}\beta)\operatorname{sen}\beta$, lo que da la ecuación

$$x\cos\beta + y\operatorname{sen}\beta - d = 0.$$

Esta es la *forma normal* de la ecuación de la recta l. Obsérvese que esta ecuación no depende del sentido positivo asignado a l, pues un cam-

bio de éste haría cambiar el signo de todos los términos del primer miembro, y, en consecuencia, la ecuación quedaría invariable.

Si se multiplica la ecuación normal por un factor arbitrario, resulta la forma general de la ecuación de la recta

$$ax + by + c = 0.$$

Para obtener de nuevo, a partir de esta forma general, la forma normal —tan llena de significado geométrico— debemos multiplicar por un factor que reduzca los dos primeros coeficientes a cos β y sen β, la suma de cuyos cuadrados ha de ser 1. Esto se consigue multiplicando por el factor $1/\sqrt{a^2 + b^2}$, que nos da la forma normal

$$\frac{a}{\sqrt{a^2 + b^2}}\, x + \frac{b}{\sqrt{a^2 + b^2}}\, y + \frac{c}{\sqrt{a^2 + b^2}} = 0,$$

de modo que tenemos:

$$\frac{a}{\sqrt{a^2 + b^2}} = \cos\beta, \qquad \frac{b}{\sqrt{a^2 + b^2}} = \operatorname{sen}\beta, \qquad -\frac{c}{\sqrt{a^2 + b^2}} = d.$$

21. Demuéstrese: *a)* que los únicos factores que reducen la forma general a la normal son $1/\sqrt{a^2 + b^2}$ y $-1/\sqrt{a^2 + b^2}$; *b)* que la elección de uno u otro de estos factores determina el sentido positivo asignado a la recta, y *c)* que cuando se utiliza uno de estos factores, el origen se halla en el lado positivo o negativo de la recta orientada resultante, o bien sobre la propia recta, según que d sea negativo, positivo o nulo.

22. Demuéstrese directamente que la recta de pendiente m que pasa por un punto dado $P_0(x_0, y_0)$ está representada por la ecuación

$$y - y_0 = m(x - x_0), \quad \text{o} \quad y = mx + y_0 - mx_0\,.$$

Demuéstrese que la recta determinada por los dos puntos $P_1(x_1, y_1)$, $P_2(x_2, y_2)$ tiene la ecuación

$$(y_2 - y_1)(x - x_1) = (x_2 - x_1)(y - y_1).$$

La abscisa x del punto donde una recta o curva corta al eje de las x se llama *abscisa en el origen* de la línea; análogamente para la *ordenada en el origen.*

23. Si se divide la ecuación general del ejercicio 20 por una constante apropiada, se ve que la ecuación de una recta puede escribirse en la forma

$$\frac{x}{a} + \frac{y}{b} = 1\,,$$

donde a y b son la abscisa y la ordenada en el origen. ¿Qué excepciones se presentan?

24. Por un procedimiento análogo, pruébese que la ecuación de una recta no paralela al eje y puede escribirse en la forma

$$y = mx +$$

(Si la recta es paralela al eje y, la ecuación correspondiente es $x=a$.)

25. Sean $ax + by + c = 0$ y $a'x + b'y + c' = 0$ las ecuaciones de dos rectas l y l', de pendientes m y m', respectivamente. Demuéstrese que l y l' son paralelas o perpendiculares según que: *a)* $m = m'$ o $mm' = -1$; *b)* $ab' - a'b = 0$ o $aa' + bb' = 0$ [obsérvese que *b)* subsiste aun cuando la recta no tenga pendiente, esto es, sea paralela al eje y].

26. Demuéstrese que la ecuación de una recta que pasa por un punto dado $P_0(x_0, y_0)$ y es paralela a una recta dada l, $ax+by+c=0$, tiene la ecuación $ax + by = ax_0 + by_0$. Pruébese que subsiste una fórmula análoga, $bx - ay = bx_0 - ay_0$, para la ecuación de la recta que pasa por P_0 y es perpendicular a l. (Obsérvese que si la ecuación de l se da en la forma normal, así resulta también la nueva ecuación en cada uno de los casos.)

27. Sean $x \cos \beta + y \operatorname{sen} \beta - d = 0$ y $ax + by + c = 0$ las formas normal y general de la ecuación de una recta l. Demuéstrese que la distancia orientada h de l a un punto cualquiera $Q(u, v)$ viene dada por

$$h = u \cos \beta + v \operatorname{sen} \beta - d,$$

o por

$$h = \frac{au + bv + c}{\pm\sqrt{a^2 + b^2}};$$

y que h es positivo o negativo según que Q esté del lado positivo o negativo de la recta orientada l (el sentido ha sido determinado por β, o por la elección del signo antepuesto a $\sqrt{a^2 + b^2}$). (Escríbase la forma normal de la ecuación de la recta m que pasa por Q y es paralela a l y hállese la distancia de l a m.)

28. Representemos por $l(x, y) = 0$ la ecuación de la recta $ax + by + c = 0$; análogamente, $l'(x, y) = 0$. Sean λ y λ' constantes tales que $\lambda + \lambda' = 1$. Demuéstrese que si l y l' se cortan en $P_0(x_0, y_0)$, la ecuación de cualquier recta que pase por P_0 es de la forma

$$\lambda l(x, y) + \lambda' l'(x, y) = 0,$$

y recíprocamente; y que cada una de tales rectas queda unívocamente determinada por la elección de un par de valores para λ y λ' (P_0 está sobre l si y sólo si $l(x_0, y_0) = ax_0 + by_0 + c = 0$.) ¿Qué rectas quedan representadas cuando l y l' son paralelas? Obsérvese que la condi-

ción $\lambda + \lambda' = 1$ es innecesaria, pero sirve para determinar una sola ecuación para cada recta que pasa por P_0.

29. Utilícese el resultado del ejercicio anterior para hallar la ecuación de una recta que pase por la intersección P_0 de l y l y por otro punto, $P_1(x_1, y_1)$, sin determinar las coordenadas de P_0. (Determínense λ y λ' por las condiciones $\lambda l(x_1, y_1) + \lambda' l'(x_1, y_1) = 0, \lambda + \lambda' = 1$.) Compruébese el resultado determinando las coordenadas de P_0 (véase pág. 85) y demuéstrese que P_0 pertenece a la recta cuya ecuación acaba de escribirse.

30. Demuéstrese que las ecuaciones de las bisectrices de los ángulos formados por las rectas no paralelas l y l' son:

$$\sqrt{a'^2 + b'^2}\, l(x, y) = \pm \sqrt{a^2 + b^2}\, l'(x, y)$$

(véase ejercicio 27). ¿Qué representan estas ecuaciones si l y l' son paralelas?

31. Hállese la ecuación de la mediatriz del segmento P_1P_2, utilizando cada uno de los siguientes métodos: *a)* determínese primero la ecuación de la recta P_1P_2, después las coordenadas del punto medio P_0 del segmento P_1P_2, y, finalmente, la ecuación de la recta que pasa por P_0 y es perpendicular a P_1P_2; *b)* escríbase la ecuación que expresa que la distancia de P_1 a un punto cualquiera de la mediatriz $P(x, y)$ es igual a la distancia de P_2 a P. Basta elevar después ambos miembros al cuadrado y simplificar.

32. Hállese la ecuación de la circunferencia determinada por tres puntos no colineales P_1, P_2, P_3, aplicando cada uno de los siguientes métodos: *a)* escríbanse primero las ecuaciones de las mediatrices de los segmentos P_1P_2 y P_2P_3; a continuación determínense las coordenadas del centro, hallando la intersección de estas dos rectas, y, finalmente, hállese el radio como distancia entre el centro y P_1; *b)* la ecuación debe ser de la forma $x^2 + y^2 - 2ax - 2by = k$ (véase página 83). Como cada uno de los puntos dados pertenece a la circunferencia, debe tenerse:

$$x_1^2 + y_1^2 - 2ax_1 - 2by_1 = k;$$
$$x_2^2 + y_2^2 - 2ax_2 - 2by_2 = k;$$
$$x_3^2 + y_3^2 - 2ax_3 - 2by_3 = k,$$

ya que un punto pertenece a la curva si y sólo si sus coordenadas satisfacen a la ecuación de la misma. Basta ahora resolver este sistema de ecuaciones respecto a a, b y k.

33. Para hallar la ecuación de la elipse de eje mayor $2p$, eje menor $2q$ y focos $F(e, 0)$ y $F(-e, 0)$, siendo $e^2 = p^2 - q^2$, hágase uso

de las distancias r y r' de los focos F y F' a un punto cualquiera de la curva. Por definición de elipse, $r + r' = 2p$, y por medio de la fórmula para la distancia, dada en la página 82, véase que

$$r'^2 - r^2 = (x + e)^2 - (x - e)^2 = 4ex.$$

Como

$$r'^2 - r^2 = (r' + r)(r' - r) = 2p(r' - r),$$

se tiene que $r' - r = 2ex/p$. Resuélvase el sistema formado por esta ecuación y la $r' + r = 2p$ y determínense las importantes fórmulas

$$r = -\frac{e}{p}x + p, \qquad r' = \frac{e}{p}x + p.$$

Como (por la fórmula de la distancia) $r^2 = (x - e)^2 + y^2$, igualando esta expresión de r^2 a la que se acaba de obtener, $\left(-\frac{e}{p}x + p\right)^2$, resulta:

$$(x - e)^2 + y^2 = \left(-\frac{e}{p}x + p\right)^2.$$

Basta ahora desarrollar, agrupar términos semejantes, sustituir $p^2 - q^2$ por e^2 y simplificar. Compruébese que el resultado puede expresarse en la forma

$$\frac{x^2}{p^2} + \frac{y^2}{q^2} = 1.$$

Hágase lo mismo para la hipérbola, definida como el lugar de los puntos P para los cuales el valor absoluto de la diferencia $r - r'$ es igual a una cantidad dada $2p$. En este caso, $e^2 = p^2 + q^2$.

34. Se define la parábola como el lugar de los puntos cuya distancia a una recta fija (directriz) es igual a su distancia a un punto dado (foco). Si elegimos la recta $x = -a$ como directriz y el punto $F(a, 0)$ como foco, compruébese que la ecuación de la parábola puede escribirse en la forma $y^2 = 4ax$.

Construcciones geométricas.

35. Demuéstrese la imposibilidad de construir con la regla y el compás los números $\sqrt[3]{3}$, $\sqrt[3]{4}$, $\sqrt[3]{5}$. Pruébese que la construcción de $\sqrt[3]{a}$ sólo es posible si a es el cubo de un número racional (véanse páginas 146 y siguientes).

36. Hállense los lados de los polígonos regulares de $3 \cdot 2^n$ y $5 \cdot 2^n$ lados y caracterícense las correspondientes sucesiones de extensiones del campo.

37. Demuéstrese la imposibilidad de trisecar con regla y compás los ángulos de 120^o y 30^o. (En el caso de 30^o, la ecuación de que de-

pende la demostración es $4z^3 - 3z = \cos 30^\circ = {}^1/_2 \sqrt{3}$. Introdúzcase como nueva incógnita $u = z\sqrt{3}$ y obténgase una ecuación en z a partir de la cual se deduce la imposibilidad de la construcción de z, en igual forma que en el texto, pág. 151.)

38. Demuéstrese que el eneágono regular no es construíble.

39. Demuéstrese que la inversión de un punto $P(x, y)$ en el punto $P'(x', y')$, respecto al círculo de radio r y centro en el origen, está definida por las ecuaciones

$$x' = \frac{xr}{x^2 + y^2}, \qquad y' = \frac{yr}{x^2 + y^2}$$

Hállense algebraicamente las ecuaciones que definen x e y en función de x' e y'.

*40. Demuéstrese analíticamente, utilizando el resultado anterior, que el conjunto de todas las circunferencias y rectas del plano se transforma en sí mismo por inversión. Compruébense las propiedades *a)* hasta *d)* de la página 155 por separado, y en forma análoga las transformaciones correspondientes a la figura 61.

41. ¿En qué se transforman las dos familias de rectas x=constante e y = constante mediante una inversión respecto a una circunferencia de centro en el origen y radio 1? Hállese la respuesta sin geometría analítica y también haciendo uso de ésta (véase pág. 172).

42. Efectúense las construcciones de Apolonio para los casos más sencillos que elegirá el lector. Inténtese la solución analítica siguiendo el método de la página 136.

Geometría proyectiva y geometría no euclídea.

43. Determínense todos los valores de la razón doble λ de cuatro puntos armónicos al efectuar todas las permutaciones de éstos. (Respuesta: $\lambda = -1, 2, {}^1/_2$.)

44. ¿Para qué configuraciones de cuatro puntos coinciden algunos de los seis valores de la razón doble dados en la página 188? (Respuesta: sólo para $\lambda = -1$ y $\lambda = 1$; existe también un valor imaginario de λ para el cual $\lambda = 1/(1 - \lambda)$, razón doble «equianarmónica».)

45. Demuéstrese que si la razón doble $(ABCD)$ vale 1, coinciden los puntos C y D.

46. Demuéstrense las proposiciones dadas en la página 188 respecto a la razón doble de cuatro planos.

47. Demuéstrese que si P y P' son inversos respecto a una circunferencia y si el diámetro AB es colineal con P y P', los puntos A, B, P, P', forman una cuaterna armónica. (Utilícese la expresión [2]

de la pág. 189, tomando como circunferencia el círculo unidad y AB como eje.)

48. Hállense las coordenadas del cuarto punto armónico de tres puntos P_1, P_2, P_3. ¿Qué ocurre si P_3 pasa a ser el punto medio de P_1P_2 (véase pág. 190).

*49. Hágase uso de las esferas de Dandelin para desarrollar la teoría de las cónicas. Demuéstrese en particular que todas son (salvo la circunferencia) lugares geométricos de puntos cuyas distancias a un punto fijo F y a una recta dada l están en una razón constante k. Para $k > 1$ se tiene la hipérbola; para $k = 1$, la parábola, y para $k < 1$, la elipse. La recta l se obtiene como intersección del plano de la cónica con el plano de la circunferencia de contacto de la esfera de Dandelin y el cono. (Ya que la circunferencia no queda incluída en esta caracterización si no es como caso límite, no resulta completamente adecuado elegir esta propiedad para definir las cónicas, si bien se hace así algunas veces.)

50. Estúdiese la proposición siguiente: «una cónica considerada a la vez como conjunto de puntos y como conjunto de rectas es una figura dual de sí misma» (véase pág. 221).

*51. Inténtese demostrar el teorema de Desargues en el plano por medio de un paso al límite efectuado en la configuración tridimensional de la figura 73 (véase pág. 184).

*52. ¿Cuántas rectas pueden trazarse que se apoyen en cuatro rectas dadas del espacio? ¿Cómo pueden ser caracterizadas? (Trácese una hiperboloide por tres de las rectas dadas; véase pág. 224.)

*53. Si el círculo de Poincaré es el círculo unidad del plano complejo, dos puntos z_1 y z_2 y los valores w_1 y w_2 de los dos puntos de intersección de la «recta» que pasa por estos dos puntos con la circunferencia unidad, definen una razón doble $\frac{z_1 - w_1}{z_1 - w_2} : \frac{z_2 - w_1}{z_2 - w_2}$ que, de acuerdo con lo dicho en el ejercicio 8 de la página 107, es real. Por definición, su logaritmo es la distancia hiperbólica entre z_1 y z_2.

*54. Transfórmese mediante una inversión el círculo de Poincaré en el semiplano superior. Desarróllese el modelo de Poincaré y sus propiedades para este semiplano, directamente y mediante esta inversión (véase pág. 236).

Topología.

55. Verifíquese la fórmula de Euler para los cinco poliedros regulares, así como para otros poliedros, llevando a cabo las correspondientes reducciones de la red de polígonos.

56. En la demostración de la fórmula de Euler (pág. 251) se trataba de reducir una red plana de triángulos, por sucesiva aplicación de dos operaciones fundamentales, en otra red formada por un solo triángulo, para la cual $V - A + C = 3 - 3 + 1 = 1$. ¿Cómo podremos asegurarnos de que el resultado final no será una *pareja* de triángulos sin ningún vértice común, de forma que $V-A+C=6-6+2=2$? (Se supone que la red primitiva es *conexa*; esto es, que se puede pasar de un vértice a otro a lo largo de los lados de la red. Demuéstrese que esta propiedad se conserva en las dos operaciones fundamentales.)

57. Hemos admitido sólo dos operaciones fundamentales para la reducción de la red antedicha; ¿no puede suceder que en algún paso un triángulo tenga sólo un vértice común con otros triángulos de la red? (Constrúyase un ejemplo.) Esto requeriría una tercera operación: supresión de dos vértices, tres aristas y una cara; ¿afectaría esto a la demostración?

58. ¿Se puede arrollar una tira ancha de caucho tres veces alrededor de un palo de escoba de forma que quede plana (esto es, no retorcida) sobre el palo? (Naturalmente, la tira de caucho debe cruzarse consigo misma en algún punto.)

59. Demuéstrese que un disco circular del cual se ha separado un punto en el centro, admite una transformación continua sin punto fijo, en sí mismo.

*60. La transformación que traslada cada punto de un disco una unidad en una dirección determinada, carece evidentemente de puntos fijos. Por supuesto, no es ésta una transformación del disco en *sí mismo*, ya que algunos puntos quedan exteriores al disco. ¿Por qué no es válido el razonamiento de la página 267 basado en la transformación $P \rightarrow P^*$?

61. Supóngase un neumático de caucho cuyo interior está pintado de blanco y su exterior de negro. ¿Es posible, si se efectúa un pequeño orificio, se deforma el neumático y a continuación se cierra el agujero, que la parte interior quede fuera, de modo que el interior sea negro y el exterior blanco?

*62. Demuéstrese que no existe «problema de los cuatro colores» en tres dimensiones, probando que para cualquier número dado n se pueden colocar n cuerpos en el espacio de manera que cada uno toque a los restantes.

*63. Haciendo uso de una superficie tórica (interior de un neumático, arganeo) o de una región plana con identificación de contorno (figura 143), constrúyase un mapa formado por siete regiones, de manera que cada una tenga frontera común con las restantes (véase pág. 260).

64. El tetraedro tetradimensional de la figura 118 consta de cinco puntos a, b, c, d y e, cada uno de los cuales está unido a los otros cuatro. Incluso si los segmentos que los unen pueden curvarse, no es posible dibujar la figura en el plano de forma que no se crucen dos cualesquiera de las conexiones. Otra configuración con diez conexiones que tampoco puede ser trazada en el plano sin cruces es la formada por seis puntos, a, b, c, a', b', c', tales que cada uno de los puntos a, b, c, esté unido a cada uno de los a', b', c'. Verifíquense experimentalmente estos hechos e inténtese una demostración basada en el teorema de la curva de Jordan. (Ha sido demostrado que cualquier configuración de puntos y rectas que no pueda representarse en el plano sin cruces debe contener como parte una de estas dos configuraciones.)

65. Una configuración está formada por las seis aristas de un tetraedro tridimensional a las que se ha añadido el segmento que une los puntos medios de dos aristas opuestas. (Dos aristas de un tetraedro son opuestas si carecen de vértice común.) Pruébese que esta configuración es equivalente a una de las descritas en el ejercicio precedente.

*66. Sean p, q, r, las tres puntas del símbolo E. El símbolo se desplaza una cierta distancia, lo que da lugar a otra E de puntas p', q', r'. ¿Puede unirse p con p', q con q' y r con r' por tres curvas que no se crucen entre sí ni tampoco a los dos símbolos?

Si recorremos el contorno de un cuadrado cambiamos nuestra dirección cuatro veces, cada una 90°, lo que hace un total de $\Delta = 360°$. Si recorremos el contorno de un triángulo, es sabido por geometría elemental que $\Delta = 360°$.

67. Demuéstrese que si C es un polígono simple cerrado, $\Delta = 360°$. (Descompóngase el interior de C en triángulos y sepárense segmentos del contorno como en la pág. 251. Sean los sucesivos contornos B_1, B_2, B_3, ... B_n. Entonces, $B_1 = C$ y B_n es un triángulo. Pruébese que si Δ_i corresponde a B_i, se verifica que $\Delta_i = \Delta_{i-1}$.)

*68. Sea C una curva simple cerrada con vector tangente que gira con continuidad. Si Δ representa la variación total del ángulo de la tangente al recorrer una vez la curva, pruébese que $\Delta = 360°$. (Sean p_0, p_1, p_2, ... p_n, p_0, puntos que descomponen C en pequeños arcos aproximadamente rectilíneos. Sea C_i la curva que corresponde a los segmentos p_0p_1, p_1p_2, ..., $p_{i-1}p_i$ y los arcos originales p_ip_{i+1}, ..., p_np_0. Entonces, $C_0 = C$ y C_n se compone de segmentos rectilíneos. Demuéstrese que $\Delta_i = \Delta_{i+1}$, y utilícese el resultado del ejercicio precedente.) ¿Es aplicable esto a la hipocicloide de la figura 55?

69. Demuéstrese que si en el diagrama de la botella de Klein de la página 275 las cuatro flechas se trazan en el sentido de las agujas

del reloj, la superficie así formada es equivalente a una esfera con un disco reemplazado por una cofia cruzada. (Esta superficie es topológicamente equivalente al plano de la geometría proyectiva.)

70. La botella de Klein de la figura 142 puede cortarse en dos mitades simétricas por medio de un plano. Demuéstrese que el resultado son dos cintas de Moebius.

*71. En la cinta de Moebius de la figura 139 se identifican los dos extremos de cada segmento transversal; demuéstrese que el resultado equivale topológicamente a una botella de Klein.

Todas las posibles parejas de puntos de un segmento rectilíneo (cuyos extremos coinciden o no) forman un cuadrado en el sentido siguiente: Si se designan los puntos del segmento por sus distancias x, y a un extremo A, las parejas ordenadas de números (x, y) pueden considerarse como las coordenadas cartesianas de un punto del cuadrado.

Todas las posibles parejas de puntos sin tener en cuenta el orden [es decir, $(x, y) = (y, x)$] forman una superficie S que equivale topológicamente a un cuadrado. Para verlo, elíjase aquella representación que tiene el primer punto más próximo al extremo A del segmento si $x \neq y$. Por tanto, S es el conjunto de todos los pares (x, y) en los cuales o x es menor que y o $x = y$. Utilizando coordenadas cartesianas, esto da el triángulo plano de vértices (0,0), (0,1), (1,1).

*72. ¿Qué superficie se forma considerando el conjunto de todos los pares ordenados de puntos de los cuales el primero pertenece a una recta y el segundo a una circunferencia? (Respuesta: un cilindro.)

73. ¿Qué superficie se forma al considerar el conjunto de todos los pares ordenados de puntos de un círculo? (Respuesta: un toro.)

*74. ¿Qué superficie resulta al considerar el conjunto de todos los pares *no ordenados* de puntos de un círculo? (Respuesta: una cinta de Moebius.)

75. He aquí las reglas de un juego para realizarlo con monedas en una gran mesa circular: A y B, por turno, colocan monedas sobre la mesa, las cuales no es necesario que se toquen y pueden colocarse en cualquier punto de la mesa en tanto que no sobresalgan del borde o cubran parte de una moneda ya colocada. Una vez colocada una moneda, no puede moverse. Con el tiempo, la mesa quedará cubierta de monedas de forma que no habrá lugar para otra, ganando aquel jugador que acierte a colocar la última moneda. Si comienza A, demuéstrese que cualquiera que sea la forma de jugar B, A puede estar seguro de ganar con tal que juegue correctamente.

76. Si en el juego del ejercicio anterior la mesa tiene la forma indicada en la figura 125 b, demuéstrese que B puede ganar siempre.

Funciones, límites y continuidad.

77. Hállese el desarrollo en fracción continua del cociente $OB : AB$ de la página 134.

78. Pruébese que la sucesión $a_0 = \sqrt{2}$, $a_{n+1} = \sqrt{2 + a_n}$ es monótona creciente y acotada por $B = 2$; en consecuencia, tiene límite. Pruébese que este límite es el número 2 (véanse págs. 135 y 337).

*79. Inténtese demostrar por métodos análogos a los utilizados en las páginas 329 y siguientes que, dada una curva cerrada sin puntos singulares (óvalo), se puede siempre trazar un cuadrado cuyos lados sean tangentes a la misma.

La función $u = f(x)$ se llama *convexa* si el punto medio del segmento que une dos puntos cualesquiera de la gráfica de la función queda por debajo de la curva; p. ej., $u = e^x$ (Fig. 278) es convexa, pero no lo es $u = \log x$ (Fig. 277).

80. Demuéstrese que la función $u = f(x)$ es convexa si y sólo si

$$\frac{f(x_1) + f(x_2)}{2} \geqslant f\left(\frac{x_1 + x_2}{2}\right).$$

El signo de igualdad únicamente se verifica para $x_1 = x_2$.

*81. Demuéstrese que para las funciones convexas se verifica la desigualdad más general

$$\lambda_1 f(x_1) + \lambda_2 f(x_2) \geqslant f(\lambda_1 x_1 + \lambda_2 x_2),$$

en la cual λ_1, λ_2, son dos constantes tales que $\lambda_1 + \lambda_2 = 1$ y $\lambda_1 \geqslant 0$, $\lambda_2 \geqslant 0$. Esto equivale a decir que ningún punto del segmento que une dos puntos de la gráfica queda por debajo de la curva.

82. Haciendo uso de la condición del ejercicio anterior, demuéstrese que las funciones $u = \sqrt{1 + x^2}$ y $u = 1/x$ (para $x > 0$) son convexas; esto es, que

$$\frac{\sqrt{1 + x_1^2} + \sqrt{1 + x_2^2}}{2} \geqslant \sqrt{1 + \left(\frac{x_1 + x_2}{2}\right)^2},$$

$$\frac{1}{2}\left(\frac{1}{x_1} + \frac{1}{x_2}\right) \geqslant \frac{2}{x_1 + x_2} \quad \text{para } x_1 \text{ y } x_2 \text{ positivos.}$$

83. Hágase lo mismo con $u = x^2$, $u = x^n$ para $x > 0$, $u = \operatorname{sen} x$ para $\pi \leqslant x \leqslant 2\pi$, $u = \operatorname{tg} x$ para $0 \leqslant x \leqslant \pi/2$, $u = -\sqrt{1 - x^2}$ para $|x| \leqslant 1$.

Máximos y mínimos.

84. Hállese el camino más corto entre P y Q como en la figura 178, en el supuesto de que la trayectoria encuentre a las dos rectas dadas n veces alternativamente (véase pág. 344).

85. Hállese el camino más corto entre dos puntos P y Q interiores a un triángulo acutángulo si la trayectoria debe encontrar a los lados del triángulo en un orden dado (véase pág. 345).

86. Trácense las curvas de nivel y compruébese la existencia de dos puntos de ensilladura, por lo menos, en una superficie sobre un dominio triplemente conexo cuya frontera se halla al mismo nivel (véase pág. 355). Se excluye el caso en que el plano tangente a la superficie es horizontal a lo largo de una curva cerrada.

87. Partiendo de dos números racionales positivos arbitrarios a, b, fórmense paso a paso los pares de números $a_{n+1}=\sqrt{a_n b_n}$, $b_{n+1} = {}^1/_2\,(a_n + b_n)$. Demuéstrese que definen un encaje de intervalos. (El punto límite cuando $n \to \infty$, que recibe el nombre de media aritmético-geométrica de a_0 y b_0, desempeñó un papel importante en las primeras investigaciones de Gauss.)

88. Hállese la longitud de toda la curva de la figura 219, y compárese con la longitud total de las dos diagonales.

*89. Hállense las condiciones que determinan si cuatro puntos A_1, A_2, A_3, A_4, se encuentran en el caso de la figura 216 o de la 218.

*90. Determínense sistemas de cinco puntos para los cuales existan diferentes redes de caminos que satisfagan las condiciones angulares. Sólo algunas de ellas serán mínimos relativos (véase pág. 370).

91. Establézcase la desigualdad de Schwarz

$$(a_1b_1 + \cdots + a_nb_n)^2 < (a_1^2 + \cdots + a_n^2)\,(b_1^2 + \cdots + b_n^2),$$

válida para cualquier conjunto de pares de números a_i, b_i; demuéstrese que el signo de igualdad solamente se verifica si los a_i son proporcionales a los b_i. (Generalícese la fórmula algebraica del ejercicio 8 anterior.)

*92. Con n números positivos x_1, ..., x_n, formamos las expresiones s_k definidas por

$$s_k = (x_1x_2 \cdots x_k + \cdots)/C_k^n,$$

donde el símbolo «$+ \ldots$» significa que deben sumarse todos los pro-

ductos de k de estos factores, elegidos de todas las maneras posibles, cuyo número es C^n_k. Demuéstrese que

$$\sqrt[k+1]{s_{k+1}} \leqslant \sqrt[k]{s_k},$$

donde el signo de igualdad solamente se verifica si todas las x_i son iguales.

93. Para $n = 3$ estas desigualdades expresan que para tres números positivos a, b, c,

$$\sqrt[3]{abc} \leqslant \sqrt{\frac{ab + ac + bc}{3}} \leqslant \frac{a + b + c}{3}$$

¿Qué propiedades extremales del cubo están implícitas en estas desigualdades?

*94. Determínese un arco de curva de longitud mínima que una dos puntos A, B, e incluya, con el segmento AB, un área dada. (Respuesta: el arco es circular.)

95. Dados dos segmentos AB y $A'B'$, hállese un arco que una A con B, y otro que una A' con B', tales que ambos arcos incluyan, con los dos segmentos, un área dada, y al mismo tiempo tengan longitud total mínima. (Respuesta: los dos arcos son circulares y del mismo radio.)

*96. La misma cuestión cuando el número de segmentos dados es cualquiera, AB, $A'B'$, etc.

*97. Sobre dos rectas que se cortan en O, determínense dos puntos A y B, respectivamente, y únanse mediante un arco de longitud mínima y tal que el área comprendida por éste y las rectas sea dada. (Respuesta: el arco es circular y perpendicular a ambas rectas.)

*98. El mismo problema, pero incluyendo ahora el perímetro total del dominio; es decir, el arco, más OA, más OB, ha de ser mínimo. (Respuesta: la solución es un arco de circunferencia con la concavidad hacia afuera y tangente a ambas rectas.)

*99. La misma cuestión para varios sectores angulares.

*100. Demuéstrese que las superficies casi planas de la figura 240 no son planas salvo para la superficie estabilizadora del centro. Observación: hallar o caracterizar analíticamente estas superficies es un problema sin resolver. Lo mismo puede decirse respecto a las superficies de la figura 251. En la figura 258 se tienen en realidad doce planos de simetría que se cortan bajo ángulos de 120° en las diagonales.

Háganse algunos experimentos adicionales con películas jabonosas. Efectúense los experimentos indicados en las figuras 256 y 257, con más de tres barras. Estúdiense los casos límites al tender a cero el

volumen de aire incluído. Háganse experimentos con planos no paralelos u otras superficies distintas. Hágase que la burbuja cúbica de la figura 258 llene todo el cubo y resalte sobre las aristas, sacando de nuevo el aire e invirtiendo el proceso.

*101. Determínense dos triángulos equiláteros de perímetro total dado y área mínima. (Los triángulos deben ser congruentes. Hágase uso del cálculo.)

*102. Hállense dos triángulos de perímetro total dado y área máxima. (Respuesta: un triángulo degenera en un punto; el otro es equilátero.)

*103. Hállense dos triángulos de área total dada y perímetro mínimo.

*104. Hállense dos triángulos equiláteros de área total dada y perímetro máximo.

Cálculo.

105. Derívense las funciones $\sqrt{1+x}$, $\sqrt{1+x^2}$, $\sqrt{\frac{x+1}{x-1}}$ por aplicación directa de la definición de derivada, formando y operando con el cociente de diferencias hasta obtener el límite por sustitución de $x_1 = x$ (véase pág. 431).

106. Demuéstrese que la función $y = e^{-1/x^2}$, con $y = 0$ para $x = 0$, tiene nulas todas sus derivadas en $x = 0$.

107. Demuéstrese que la función del ejercicio anterior no puede desarrollarse en serie de Taylor (véase pág. 487).

108. Determínense los puntos de inflexión $[f''(x) = 0]$ de las curvas $y = e^{-x^2}$ e $y = x\, e^{-x^2}$.

109. Demuéstrese que si $f(x)$ es un polinomio cuyos n ceros $x_1, \ldots, x_n$, son distintos, se tiene:

$$\frac{f'(x)}{f(x)} = \sum_{i=1}^{n} \frac{1}{x - x_i}$$

*110. Utilizando la definición de integral como límite de una suma, demuéstrese que para $n \to \infty$ se tiene:

$$n\left(\frac{1}{1^2+n^2} + \frac{1}{2^2+n^2} + \cdots + \frac{1}{n^2+n^2}\right) \to \frac{\pi}{4}$$

*111. Demuéstrese en forma análoga que

$$\frac{b}{n}\left(\operatorname{sen}\frac{b}{n} + \operatorname{sen}\frac{2b}{n} + \cdots + \operatorname{sen}\frac{nb}{n}\right) \to \cos b - 1.$$

112. Dibujando la figura 276 a gran escala sobre papel milimetrado y contando los cuadraditos del área rayada, determínese un valor aproximado de π.

113. Hágase uso de la fórmula [7], página 451, para el cálculo numérico de π con error menor que $1/100$.

114. Demuéstrese que $e^{\pi i} = -1$ (véase pág. 488).

115. Una curva de forma dada se dilata en la relación $1 : x$. $L(x)$ y $A(x)$ representan la longitud y el área de la curva dilatada. Demuéstrese que $L(x)/A(x) \to 0$ cuando $x \to \infty$, y, en general, $L(x)/A(x)^k \to 0$ para x tendiendo a infinito, si $k > 1/2$. Compruébese para el círculo, el cuadrado y la *elipse. (El área tiene orden superior de magnitud que la circunferencia. Véase pág. 482.)

116. Con frecuencia la función exponencial aparece en combinaciones para las que se emplea la notación siguiente:

$$u = \operatorname{sh} x = {}^1/_2(e^x - e^{-x}), \qquad v = \operatorname{ch} x = {}^1/_2(e^x + e^{-x})$$

$$w = \operatorname{th} x = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

y que se llaman *seno hiperbólico, coseno hiperbólico* y *tangente hiperbólica*, respectivamente. Estas funciones gozan de muchas propiedades análogas a las de las funciones trigonométricas y se hallan relacionadas con la hipérbola $u^2 - v^2 = 1$ en forma análoga a como las funciones $u = \cos x$ y $v = \operatorname{sen} x$ están relacionadas con la circunferencia $u^2 + v^2 = 1$. El lector establecerá las siguientes propiedades, comparándolas con las correspondientes de las funciones trigonométricas:

$$D \operatorname{ch} x = \operatorname{sh} x, \qquad D \operatorname{sh} x = \operatorname{ch} x, \qquad D \operatorname{th} x = 1/\operatorname{ch}^2 x,$$
$$\operatorname{sh}(x + x') = \operatorname{sh} x \cdot \operatorname{ch} x' + \operatorname{ch} x \operatorname{sh} x',$$
$$\operatorname{ch}(x + x') = \operatorname{ch} x \cdot \operatorname{ch} x' + \operatorname{sh} x \cdot \operatorname{sh} x'.$$

Las funciones inversas son: $x = \arg \operatorname{sh} u = \log(u + \sqrt{u^2 + 1})$; $x = \arg \operatorname{ch} v = \log(v + \sqrt{v^2 - 1})$ $(v \geqslant 1)$.

Sus derivadas son:

$$D \arg \operatorname{sh} u = \frac{1}{\sqrt{1 + u^2}}; \qquad D \arg \operatorname{ch} v = \frac{1}{\sqrt{v^2 - 1}}$$

$$D \arg \operatorname{th} w = \frac{1}{1 - w^2}, \qquad (|w| > 1).$$

117. Hágase uso de la fórmula de Euler para comprobar la analogía de las funciones hiperbólicas y trigonométricas.

*118. Hállense fórmulas sencillas de sumación para

$$\operatorname{sh} x + \operatorname{sh} 2x + \cdots + \operatorname{sh} nx$$

y

$$1/2 + \operatorname{ch} x + \operatorname{ch} 2x + \cdots + \operatorname{ch} nx$$

análogas a las dadas en el ejercicio 14 para las funciones trigonométricas.

Técnica de la integración.

El teorema de la página 449 reduce el problema de integrar una función $f(x)$, entre los límites a y b, al de hallar una función primitiva $G(x)$ de $f(x)$; es decir, tal que $G'(x) = f(x)$. La integral es entonces la diferencia $G(b) - G(a)$. Es habitual utilizar para estas funciones primitivas, determinadas por $f(x)$ (salvo una constante aditiva arbitraria), el nombre de «integral indefinida» y la sugerente notación

$$G(x) = \int f(x)\,dx.$$

(Esta notación puede resultar confusa para el principiante; véase la observación de la pág. 448.)

Toda fórmula de derivación contiene la solución de un problema de integración indefinida sin más que interpretarla al revés como fórmula de integración. Podemos extender este procedimiento un tanto empírico mediante dos importantes reglas que no son sino las equivalentes a las reglas de derivación de una función compuesta y de un producto de funciones. En su forma integral, dichas reglas se llaman de *integración por sustitución* y de *integración por partes.*

A) La primera regla resulta de la fórmula de derivación de una función compuesta

$$H(u) = G(x),$$

en la que

$$x = \psi(u) \qquad \text{y} \qquad u = \varphi(x)$$

se suponen funciones inversas, unívocamente determinadas en el intervalo que se considera. Se tiene entonces

$$H'(u) = G'(x)\psi'(u).$$

Si

$$G'(x) = f(x),$$

podemos escribir

$$G(x) = \int f(x)\,dx$$

y también

$$G'(x)\psi'(u) = f(x)\psi'(u),$$

lo cual, como consecuencia de la fórmula anterior para $H'(u)$, equivale a

$$H(u) = \int f(\psi(u))\psi'(u)\,du.$$

Por tanto, ya que $H(u) = G(x)$,

$$\int f(x)\,dx = \int f(\psi(u))\psi'(u)\,du. \qquad [I]$$

Escrita con la notación de Leibniz (véase pág. 443), esta regla adquiere esta otra forma:

$$\int f(x)\,dx = \int f(x)\frac{dx}{du}\,du,$$

que significa que el símbolo dx debe sustituirse por el símbolo $\frac{dx}{du}\,du$, exactamente como si dx y du fueran números y $\frac{dx}{du}$ una fracción.

La utilidad de la fórmula [I] queda aclarada con unos cuantos ejemplos:

a) $J = \int \frac{1}{u \log u}\,du$. Comenzamos aquí con el segundo miembro de [1], sustituyendo $x = \log u = \psi(u)$. Tenemos así $\psi'(u) = 1/u$, $f(x) = 1/x$; por tanto,

$$J = \int \frac{dx}{x} = \log x,$$

o

$$\int \frac{du}{u \log u} = \log \log u.$$

Podemos comprobar este resultado derivando ambos miembros, lo que nos da $\frac{1}{u \log u} = \frac{d}{du}(\log \log u)$, que, como fácilmente se comprueba, es cierto.

b) $J = \int \cot u\,du = \int \frac{\cos u}{\operatorname{sen} u}\,du$. Haciendo $x = \operatorname{sen} u = \psi(u)$ obtenemos:

$$\psi'(u) = \cos u, \qquad f(x) = x;$$

por tanto,

$$J = \int \frac{dx}{x} = \log x$$

o

$$\int \cot u \, du = \log \operatorname{sen} u \, .$$

Este resultado puede también comprobarse por derivación.

c) En general, si tenemos una integral de la forma

$$J = \int \frac{\psi'(u)}{\psi(u)} \, du,$$

hacemos $x = \psi(u)$, $f(x) = x$, y se obtiene:

$$J = \int \frac{dx}{x} = \log x = \log \psi(u).$$

d) $J = \int \operatorname{sen} x \cos x \, dx$. Hacemos $\operatorname{sen} x = u$, $\cos x = \dfrac{du}{dx}$, y se tiene:

$$J = \int u \frac{du}{dx} \, dx = \int u \, du = \frac{u^2}{2} = {}^1/_2 \operatorname{sen}^2 x.$$

e) $J = \int \dfrac{\log u}{u} \, du$. Hacemos $\log u = x$, $\dfrac{1}{u} = \dfrac{dx}{du}$, y obtenemos:

$$J = \int x \frac{dx}{du} \, du = \int x \, dx = \frac{x^2}{2} = {}^1/_2 (\log u)^2.$$

En los ejemplos que siguen se utiliza [I] comenzando con el primer miembro.

f) $J = \int \dfrac{dx}{\sqrt{x}}$. Si se hace $\sqrt{x} = u$, se tiene: $x = u^2$ y $\dfrac{dx}{du} = 2u$; por tanto,

$$J = \int \frac{1}{u} \cdot 2u \, du = 2u = 2\sqrt{x}\, .$$

g) Mediante la sustitución $x = au$, siendo a constante, obtenemos:

$$\int \frac{dx}{a^2 + x^2} = \int \frac{dx}{du} \cdot \frac{1}{a^2} \cdot \frac{1}{1 + u^2} \, du = \int \frac{1}{a} \frac{du}{1 + u^2} = \frac{1}{a} \cdot \operatorname{arc\,tg} \frac{x}{a}$$

h) $J = \int \sqrt{1 - x^2}\, dx$. Hagamos $x = \cos u$, $\dfrac{dx}{du} = -\text{sen}\, u$, con lo que

$$J = -\int \text{sen}^2 u\, du = -\int \frac{1 - \cos 2u}{2}\, du = -\frac{u}{2} + \frac{\text{sen}\, 2u}{4}$$

Si utilizamos la fórmula $\text{sen}\, 2u = 2\, \text{sen}\, u \cos u = 2 \cos u \sqrt{1 - \cos^2 u}$, obtenemos:

$$J = -\tfrac{1}{2} \arccos x + \tfrac{1}{2} x \sqrt{1 - x^2}.$$

Determínense las integrales indefinidas siguientes, comprobando los resultados por derivación:

119) $\displaystyle\int \frac{u\, du}{u^2 - u + 1}$

120) $\displaystyle\int u e^{u^2}\, du.$

121) $\displaystyle\int \frac{du}{u(\log u)^n}$

122) $\displaystyle\int \frac{8x}{3 + 4x}\, dx.$

123) $\displaystyle\int \frac{dx}{x^2 + x + 1}$

124) $\displaystyle\int \frac{dx}{x^2 + 2ax + b}$

125) $\displaystyle\int t^2 \sqrt{1 + t^3}\, dt.$

126) $\displaystyle\int \frac{t + 1}{\sqrt{1 - t^2}}\, dt.$

127) $\displaystyle\int \frac{t^4}{1 - t}\, dt.$

128) $\displaystyle\int \cos^n t \cdot \text{sen}\, t \cdot dt.$

129) Demuéstrese que

$$\int \frac{dx}{a^2 - x^2} = \frac{1}{a} \operatorname{arg\,th} \frac{x}{a}\,; \quad \int \frac{dx}{\sqrt{a^2 - x^2}} = \operatorname{arg\,sh} \frac{x}{a}$$

[Compárese con los ejemplos *g)* y *h)*].

B) La regla (pág. 437) para derivar un producto,

$$[p(x) \cdot q(x)]' = p(x) \cdot q'(x) + p'(x) \cdot q(x),$$

puede escribirse como fórmula integral

$$p(x) \cdot q(x) = \int p(x) \cdot q'(x)\, dx + \int p'(x) \cdot q(x)\, dx$$

o

$$\int p(x) \cdot q'(x)\, dx = p(x)q(x) - \int p'(x) \cdot q(x)\, dx. \qquad \text{[II]}$$

Escrita en esta forma recibe el nombre de regla de *integración por partes*. Esta fórmula es muy útil cuando la función que ha de inte-

grarse puede escribirse en la forma $p(x)\, q'(x)$, donde se conoce la función primitiva $q(x)$ de $q'(x)$. En tal caso, la fórmula [II] reduce el problema de hallar la integral indefinida de $p(x)\, q'(x)$ al de la integración de la función $p'(x)\, q(x)$, que es con frecuencia más fácil.

Ejemplos:

a) $J = \int \log x \, dx$. Hagamos $p(x) = \log x$, $q'(x) = 1$; de forma que $q(x) = x$. Aplicando [II] se obtiene:

$$\int \log x \, dx = x \log x - \int \frac{x}{x} \, dx = x \log x - x.$$

b) $J = \int x \log x \, dx$. Hagamos $p(x) = \log x$, $q'(x) = x$, con lo que resulta:

$$J = \frac{x^2}{2} \log x - \int \frac{x^2}{2x} \, dx = \frac{x^2}{2} \log x - \frac{x^2}{4}$$

c) $J = \int x \operatorname{sen} x \, dx$. En este caso hacemos $p(x) = x$, $q(x) = -\cos x$, y obtenemos:

$$\int x \operatorname{sen} x \, dx = -x \cos x + \operatorname{sen} x.$$

Obténgase las siguientes integrales haciendo uso de la fórmula de integración por partes:

130) $\int x e^x \, dx.$

131) $\int x^2 \cos x \, dx.$
(Aplíquese dos veces [II])

132) $\int x^a \log x \, dx \quad (a \neq -1).$

133) $\int x^2 e^x \, dx.$
(Utilícese el ejercicio 130)

La integración por partes de la integral $\int \operatorname{sen}^m x \, dx$ conduce a una expresión notable para π como producto infinito. Para deducirla, escribamos la función $\operatorname{sen}^m x$ en la forma $\operatorname{sen}^{m-1} x \operatorname{sen} x$ e integremos por partes entre los límites 0 y $\pi/2$, con lo cual obtenemos la fórmula

$$\int_0^{\pi/2} \operatorname{sen}^m x \, dx = (m-1) \int_0^{\pi/2} \operatorname{sen}^{m-2} x \cos^2 x \, dx =$$

$$= -(m-1) \int_0^{\pi/2} \operatorname{sen}^m x \, dx + (m-1) \int_0^{\pi/2} \operatorname{sen}^{m-2} x \, dx,$$

o

$$\int_0^{\pi/2} \operatorname{sen}^m x\,dx = \frac{m-1}{m}\int_0^{\pi/2} \operatorname{sen}^{m-2} x\,dx,$$

ya que el primer término del segundo miembro de [II], pq, es igual a 0 para los valores 0 y $\pi/2$. Por aplicación reiterada de la última fórmula, obtenemos el siguiente valor de $I_m = \int_0^{\pi/2} \operatorname{sen}^m x\,dx$ (las fórmulas difieren según que m sea par o impar):

$$I_{2n} = \frac{2n-1}{2n}\cdot\frac{2n-3}{2n-2}\cdots\frac{1}{2}\cdot\frac{\pi}{2},$$

$$I_{2n+1} = \frac{2n}{2n+1}\cdot\frac{2n-2}{2n-1}\cdots\frac{2}{3}.$$

Puesto que $0 < \operatorname{sen} x < 1$ para $0 < x < \pi/2$, se obtiene $\operatorname{sen}^{2n-1} x >$ $> \operatorname{sen}^{2n} x > \operatorname{sen}^{2n+1} x$, de forma que

$$I_{2n-1} > I_{2n} > I_{2n+1}$$ (véase pág. 424)

o

$$\frac{I_{2n-1}}{I_{2n+1}} > \frac{I_{2n}}{I_{2n+1}} > 1.$$

Sustituyendo los valores calculados antes para I_{2n-1}, etc., de las últimas desigualdades obtenemos:

$$\frac{2n+1}{2n} > \frac{1\cdot 3\cdot 3\cdot 5\cdot 5\cdot 7\cdots\cdots(2n-1)(2n-1)(2n+1)}{2\cdot 2\cdot 4\cdot 4\cdot 6\cdot 6\cdots(2n)(2n)}\cdot\frac{\pi}{2} > 1.$$

Si pasamos ahora al límite para $n \to \infty$, vemos que el término del centro tiende a 1, de donde obtenemos la representación de $\pi/2$ mediante el producto de Wallis:

$$\frac{\pi}{2} = \frac{2\cdot 2\cdot 4\cdot 4\cdot 6\cdot 6\cdots 2n\cdot 2n\cdots}{1\cdot 3\cdot 3\cdot 5\cdot 5\cdot 7\cdots(2n-1)(2n-1)\cdot(2n+1)\cdots} =$$

$$= \lim \frac{2^{4n}(n!)^4}{[(2n)!]^2\,(2n+1)} \text{ para } n \to \infty.$$

IX. AVANCES RECIENTES

§1. Una fórmula para los primos

[*véase página* 49]

Se conocen ahora muchos polinomios diferentes que producen números primos, aunque poco contribuyen a nuestro conocimiento sobre tales números; en lugar de ello, se demuestra que los polinomios pueden tener propiedades muy extrañas.

En su celebrada ponencia en el Congreso Internacional de Matemáticos, de 1900, David Hilbert planteó 23 problemas cuya solución sentía que sería de la mayor importancia para el avance de las matemáticas. El décimo problema de Hilbert se plantea investigar si existe un método general (lo que ahora llamaríamos *algoritmo)* para examinar si una ecuación diofantina tiene solución. En 1970, siguiendo el trabajo anterior de Martin Davis, Hilary Putnam y Julia Robinson, el matemático ruso Yuri Matyasevich demostró que no existe tal "algoritmo de decisión". Como su método utiliza precisamente polinomios como un "lenguaje de programación" un tanto aparatoso con el cual se simulan algoritmos para computadora, los polinomios producidos son realmente enormes. James Jones descubrió un sistema explícito de ecuaciones polinomiales para el que no existe ningún algoritmo de decisión: consta de 18 ecuaciones en 33 variables con grado máximo 5^{60}.

Un resultado secundario intrigante de la demostración de Matyasevich es que existe un polinomio $p(x_1, \ldots, x_n)$ igual de complicado, en 23 variables, cuyos valores *positivos* para valores enteros de las variables son precisamente los primos. En 1976, J. P. Jones, D. Sato, H. Wada y D. Wiens publicaron un polinomio relativamente simple en 26 variables con la misma propiedad. Denotemos a las variables con $a, b, c, \ldots, x, y, z$ (es una coincidencia, pero muy útil tipográficamente, que el alfabeto en inglés tenga 26 letras). Dicho polinomio es:

$$(k+2)\{1-[wz+h+j-q]^2-[(gk+2g+k+1)(h+j)+h-z]^2$$
$$-[2n+p+q+z-e]^2-[16(k+1)^3(k+2)(n+1)^2+1-f^2]^2$$
$$-[e^3(e+2)(a+1)^2+1-o^2]^2-[(a^2-1)y^2+1-x^2]^2$$
$$-[16r^2y^4(a^2-1)+1-u^2]^2-$$

$$- [((a + u^2(u^2 - a))^2 - 1)(n + 4dy^2) + 1 - (x + cu)^2]^2 - [n + l + v - y]^2 - \\ - [(a^2 - 1)l^2 + 1 - m^2]^2 - [ai + k + 1 - l - i]^2 \\ - [p + l(a - n - 1) + b(2an + 2a - n^2 - 2n - 2) - m]^2 \\ - [q + y(a - p - 1) + s(2ap + 2a - p^2 - 2p - 2) - x]^2 \\ - [z + pl(a - p) + t(2ap - p^2 - 1) - pm]^2\}.$$

Los valores *positivos* de esta expresión para valores enteros de $a, \ldots, z$ son precisamente los primos.

Aparentemente hay una paradoja: es claro que la expresión puede descomponerse en factores, de hecho es de la forma $(k + 2)\{1 - M\}$. Sin embargo, M es una suma de cuadrados, de manera que la expresión es positiva si y sólo si $M = 0$, y su valor es entonces $k + 2$. Así que el polinomio M tiene que construirse de manera que

$$M(k, \text{otras variables}) = 0 \quad \text{si y sólo si } k + 2 \text{ es primo,}$$

lo que puede lograrse usando los métodos de Matyasevich.

Este resultado se vuelve ligeramente *menos* intrigante cuando se hace claro que en este contexto nada hay de especial acerca de los primos: pueden ser reemplazados por cualquier sucesión de números "recursivamente enumerable" —lo que significa esencialmente una sucesión infinita determinada por un sistema finito de condiciones computables— construyendo un polinomio apropiado. El avance del descubrimiento radica en que el concepto de "computabilidad" puede expresarse en el lenguaje de los polinomios, no en que la teoría de los primos pueda simplificarse introduciendo una fórmula algebraica.

§2. La conjetura de Goldbach y los primos gemelos
[*véase página* 54]

La conjetura de Goldbach de que todo número par mayor que 2 es una suma de dos primos, y la cercanamente relacionada "conjetura de los primos gemelos" de que existe una infinidad de primos p para los cuales $p + 2$ también es primo, permanecen abiertas. Sin embargo, ahora se sabe bastante más acerca de ambas cuestiones.

Uno de los métodos más poderosos para abordar algunos problemas en la teoría de los números es el análisis complejo, idea que se remonta a Euler y que fue explotada en particular por Riemann en su estudio de la función zeta $\zeta(s)$

(véase página 525). A partir de 1920, Godfrey H. Hardy y John E. Littlewood investigaron la aplicación de la teoría analítica de los números, como se le llegó a llamar, a cuestiones relacionadas con la representación de números como sumas de números de tipos especiales. En 1937, I. M. Vinográdov utilizó esos métodos para demostrar que todo número impar suficientemente grande es una suma de tres primos, lo cual mejoró su resultado de cuatro primos, citado por Courant y Robbins en la página 55, que había sido demostrado en 1934. Según expresaron Courant y Robbins, este teorema sólo se aplica a números "suficientemente grandes" —números mayores que algún valor particular n_0—, aunque la demostración de Vinográdov no especifica qué tan grande debe ser n_0. En 1956, K. G. Borodzkin llenó este hueco mostrando que $n_0 = \exp[\exp(16.038)]$ es suficiente, donde $\exp(x) = e^x$. Varios matemáticos usaron el método de Vinográdov para demostrar que "casi todos" los números pares son la suma de dos primos, es decir, que la proporción de tales números hasta algún límite n tiende a 100% conforme n tiende a infinito.

En 1919, Viggo Brun introdujo un enfoque diferente, el "método de la criba", que generaliza la criba de Eratóstenes (véase página 49), el cual usó para demostrar que todo entero par suficientemente grande es una suma de dos números, siendo cada uno de ellos un producto de a lo más nueve primos. Siguieron una serie de mejoras a este teorema por parte de varias personas. Por ejemplo, en 1937, G. Ricci demostró que todo entero par suficientemente grande es una suma de dos números, siendo uno de ellos un producto de a lo más dos primos, y el otro un producto de a lo más 366 primos. P. Kuhn usó ideas sobre combinatoria de A. A. Buchstab para demostrar que todo entero par suficientemente grande es una suma de dos números, cada uno de los cuales es producto de a lo más cuatro primos. En 1957, Wang Yuan demostró que todo entero par suficientemente grande es una suma de un primo y un producto de a lo más tres primos, bajo la suposición de que se cumple la hipótesis generalizada de Riemann.

La hipótesis clásica de Riemann, otro de los 23 problemas de Hilbert y todavía (discutiblemente) la mayor cuestión sin resolverse en todas las matemáticas, tiene que ver con la función zeta de Riemann $\zeta(s)$ cuando la variable s es compleja. Específicamente, enuncia que si $\zeta(s) = 0$ y s no es real entonces $s = \frac{1}{2} + iy$ para algún real y. Las consecuencias de demostrar este enunciado serían espectaculares: revolucionarían la teoría de los números y la geometría algebraica. Más aún, cualquier método para resolver tal problema es casi seguro que se extendería a otras variantes importantes tales como la hipótesis generalizada de Riemann, un enunciado considerablemente más fuerte del mismo tipo general. Debido a que la hipótesis de Riemann y sus generalizaciones constituyen un obstáculo tan significativo para el progreso, los especialistas en teoría de números han desarrollado el hábito de arrojar anzuelos

exploratorios al territorio que queda más allá, basando parte de su trabajo en la suposición explícita de que la hipótesis de Riemann, o alguna de sus generalizaciones, es verdadera. Una justificación de este proceder es la posibilidad de que pueda llevar a una contradicción, exponiendo entonces la hipótesis de Riemann como falsa; pero esto es mera especulación. Los especialistas en teoría de números son impacientes, no pueden esperar para ver qué hay más allá del Gran Obstáculo.

En ocasiones, una vez que tal territorio ha sido explorado, aparecen nuevas posibilidades que permiten que se prescinda de la suposición. En 1948, sin suponer la hipótesis generalizada de Riemann, Alfred Rényi demostró que todo entero par suficientemente grande es una suma de un primo y un producto de a lo más c primos para algún c fijo pero desconocido. En 1961, M. B. Barban mostró que $c = 9$ basta. En 1962, Pan Cheng Dong redujo esto a $c = 5$; poco después Barban y Pan lo redujeron, independientemente, a $c = 4$, y en 1965, Buchstab demostró el teorema para $c = 3$. Finalmente, en 1966, Chen Jing Run mejoró el método de la criba y demostró el teorema para $c = 2$. Es decir, todo entero par suficientemente grande es una suma de un primo y un producto de al menos dos primos: "primo más casi primo". Éste es el resultado más cercano, que se conoce hasta la fecha, a la conjetura de Goldbach.

Se ha abordado la conjetura de los primos gemelos con un espíritu similar. El artículo de Brun de 1919 demostró también que hay una infinidad de números p tales que p y $p + 2$ son ambos un producto de a lo más nueve primos. En correspondencia a las mejoras al resultado de Brun sobre la conjetura de Goldbach, hubo mejoras similares a su trabajo sobre la conjetura de los primos gemelos. En 1924, Rademacher redujo el número nueve de Brun a siete. Buchstab lo redujo aún más, a seis en 1930 y a cinco en 1938. En un artículo de 1957, Wang indicó: "también se han obtenido resultados análogos en el problema de los primos gemelos"; dicha afirmación —puesta en el contexto de la demostración, ese mismo año, de que todo entero par suficientemente grande es una suma de un primo y un producto de a lo más tres primos— equivale a sostener que hay una infinidad de números p tales que tanto p como $p + 2$ son un producto de a lo más tres primos. Tomando como válida la hipótesis generalizada de Riemann, Wang mostró en 1962 que existe una infinidad de primos p tales que $p + 2$ es un producto de a lo más tres primos. En 1965, sin necesidad de usar la hipótesis de Riemann, Buchstab demostró que para algún c fijo existe una infinidad de primos p tales que $p + 2$ es un producto de a lo más c primos. En un artículo de 1973, Chen demostró que $c = 2$ basta y, otra vez, es éste el resultado más cercano que se conoce a la conjetura de los primos gemelos. Parece improbable que los métodos actuales puedan lograr una mejor aproximación al resultado: se requiere de una idea genuinamente novedosa.

§3. El último teorema de Fermat

[*véase página* 66]

Uno de los avances más dramáticos desde que Courant y Robbins escribieron *¿Qué son las matemáticas?* fue la demostración del último teorema de Fermat en 1994, dada por Andrew Wiles, de la Universidad de Princeton. Recordemos que Fermat conjeturó que la ecuación

$$x^n + y^n = z^n \tag{1}$$

no tiene soluciones enteras distintas de la solución cero cuando $n \geq 3$. La demostración de Wiles es muy técnica y sólo accesible a los expertos; sin embargo, el bosquejo general de la demostración sí es comprensible. El ataque es esencialmente indirecto y usa bastante la teoría de las "curvas elípticas", las cuales están definidas por ecuaciones diofantinas de la forma

$$y^2 = ax^3 + bx^2 + cx + d \tag{2}$$

para números racionales a, b, c y d. (El adjetivo "elípticas" se deriva de relaciones con las llamadas funciones elípticas, y no se refiere a la forma de la curva.) Se sabe mucho acerca de tales ecuaciones: constituyen una de las áreas mejor y más profundamente entendidas de la teoría de números.

La ecuación de Fermat (1) puede reescribirse como $(x/z)^n + (y/z)^n = 1$, de manera que el punto $(X, Y) = (x/z, y/z)$ está en la *curva de Fermat* con ecuación

$$X^n + Y^n = 1. \tag{3}$$

Se dice que (X, Y) es un punto racional si tanto X como Y son números racionales. Entonces el último teorema de Fermat es equivalente a la afirmación de que ningún punto racional puede estar en la curva de Fermat (3) cuando $n \geq 3$. Entre 1970 y 1975, Yves Hellegouarch investigó una relación curiosa entre las curvas de Fermat (3) y las curvas elípticas (2). Jean-Pierre Serre sugirió intentar lo inverso: explotar propiedades de las curvas elípticas para demostrar resultados relacionados con el último teorema de Fermat. En 1985, Gerhard Frey dio forma concreta a esta sugerencia introduciendo lo que ahora se llama la *curva elíptica de Frey*, asociada con una presumible solución de la ecuación de Fermat. Supóngase que hay una solución no trivial $A^n + B^n = C^n$ de la ecuación de Fermat, y fórmese la curva elíptica

$$y^2 = x(x + A^n)(x - B^n), \tag{4}$$

que es precisamente la curva elíptica de Frey, la cual existe si y sólo si el último teorema de Fermat es falso. Así que para demostrar el último teorema de Fermat es suficiente demostrar que la curva de Frey (4) no puede existir; la manera de hacerlo es siguiendo el método "indirecto" de demostración (véase página 113): es decir, suponer que sí existe y deducir una contradicción, lo que implicaría que la curva de Frey no existe después de todo y que por lo tanto el último teorema de Fermat es verdadero. Frey encontró fuerte evidencia de que su curva "no debiera existir" demostrando que ésta tiene muchas propiedades notablemente extrañas cuya base es de improbable solidez. En 1986 Kenneth Ribet estableció claramente el problema demostrando que la curva de Frey no puede existir siempre y cuando la conjetura de Taniyama, un gran problema abierto en teoría de números, sea verdadera. De esta manera redujo un importante problema abierto, el último teorema de Fermat, a otro problema abierto importante. Este tipo de reducción con frecuencia resulta inútil: sólo reemplaza un problema difícil por otro más difícil; pero en este caso se dio en la veta, pues proporcionó un *contexto* para atacar el problema.

La conjetura de Taniyama es muy técnica también, pero puede explicarse haciendo referencia a un caso especial. Hay una íntima relación entre la "ecuación pitagórica" $a^2 + b^2 = c^2$, el círculo unitario y las funciones trigonométricas seno y coseno. Para encontrar esta relación, obsérvese que la ecuación pitagórica puede reescribirse en la forma $(a/c)^2 + (b/c)^2 = 1$, lo cual implica que el punto $(x, y) = (a/c, b/c)$ está en el círculo unitario, cuya ecuación es $x^2 + y^2 = 1$. Es bien sabido que las funciones trigonométricas proporcionan un modo sencillo de representar el círculo unitario. Específicamente, el teorema de Pitágoras y la definición geométrica de sen y cos implican que la ecuación

$$\cos^2\theta + \operatorname{sen}^2\theta = 1 \tag{5}$$

se cumple para cualquier ángulo θ (véase página 312). Si hacemos $x = \cos\theta$ y $y = \operatorname{sen}\theta$ entonces (5) establece que el punto (x, y) está en el círculo unitario. Resumiendo: resolver la ecuación pitagórica en enteros es equivalente a encontrar un ángulo θ tal que tanto $\cos\theta$ como $\operatorname{sen}\theta$ sean números racionales (iguales respectivamente a a/c y b/c). Como las funciones trigonométricas tienen todo tipo de propiedades agradables, esta idea es la base de una teoría realmente fructífera de la ecuación pitagórica.

La conjetura de Taniyama dice (en un contexto más bien técnico) que una idea de tipo similar puede aplicarse a cualquier curva elíptica, pero reemplazando al seno y al coseno por funciones "modulares" más sofisticadas. Así, problemas sobre curvas elípticas pueden reemplazarse por problemas sobre funciones modulares, igual que problemas sobre el círculo pueden reemplazarse por problemas sobre funciones trigonométricas.

Wiles se dio cuenta de que el acercamiento de Frey puede llevarse hasta una conclusión satisfactoria sin usar toda la fuerza de la conjetura de Taniyama. En lugar de eso, basta un caso particular, uno que se aplica a una clase de curvas elípticas conocidas como "semiestables". En un artículo de cien páginas hizo acopio de un poderoso arsenal para demostrar el caso semiestable de la conjetura de Taniyama, que lleva al siguiente teorema. Supóngase que M y N son enteros diferentes, distintos de cero y primos relativos tales que $MN(M-N)$ es divisible entre 16. Entonces la curva elíptica $y^2 = x(x + M)(x + N)$ puede parametrizarse mediante funciones modulares. De hecho, la condición de divisibilidad entre 16 implica que esta curva es semiestable, de manera que la conjetura semiestable de Taniyama establece la propiedad deseada.

Ahora aplicamos el teorema de Wiles a la curva de Frey (4) haciendo $M=A^n$ y $N = -B^n$. Entonces $M - N = A^n + B^n = C^n$, de manera que $MN(M - N) = -A^nB^nC^n$, lo cual debemos mostrar que es un múltiplo de 16. Ahora, al menos uno de los números A, B y C debe ser par —pues si A y B son ambos impares, C^n es una suma de dos impares y por lo tanto par, lo cual implica que C es par—. Además podemos suponer que $n \geq 5$, pues hace mucho tiempo Euler demostró el último teorema de Fermat para $n = 3$. Pero dado que la quinta potencia o una mayor de un número par es divisible entre $2^5 = 32$, el número $-A^nB^nC^n$ es un múltiplo de 32, así que claramente también un múltiplo de 16. Por lo tanto la curva de Frey satisface la hipótesis del teorema de Wiles, implicando que puede ser parametrizada con funciones modulares. Sin embargo, la demostración de Ribet de que la conjetura de Taniyama implica la no existencia de la curva de Frey funciona demostrando que la curva de Frey *no puede* ser parametrizada con funciones modulares, lo que constituye una contradicción, así que el último teorema de Fermat es verdadero.

Esta demostración es muy indirecta y requiere ideas sofisticadas. Más aún, surgieron algunas dificultades referentes a la primera versión de la demostración de Wiles que aumentaron el dramatismo. Él hizo circular un mensaje por correo electrónico a la comunidad matemática reconociendo esas dificultades pero asegurando su confianza en que sus métodos las resolverían. Las reparaciones a la demostración tomaron más tiempo del que se esperaba, pero el 26 de octubre de 1994 Karl Rubin hizo circular otro mensaje: "Como la mayoría de ustedes sabe, el argumento descrito por Wiles [...] resultó tener un hueco importante, a saber, la construcción de un sistema de Euler. Después de tratar sin éxito de reparar esa construcción, Wiles regresó a un camino distinto, el cual ya había intentado antes pero que había abandonado en favor de la idea del sistema de Euler. Entonces pudo completar su demostración."

§4. La hipótesis del continuo

[*véase página* 115]

La hipótesis del continuo establece que el cardinal del conjunto de todos los números reales es el cardinal infinito más pequeño que es mayor que el de los enteros. Ahora se sabe que la hipótesis del continuo no es ni verdadera ni falsa, sino *indecidible*. Para entender lo que esto significa, debemos recordar brevemente el método axiomático (página 249). Dicho método particulariza un objeto matemático estableciendo un sistema explícito de condiciones, *axiomas*, que se requiere que el objeto satisfaga. Esto enfoca la atención en las relaciones abstractas entre ese objeto y otros, más que en los materiales primarios con los que está "construido". Las presentaciones sencillas de la teoría de conjuntos dan por hecho que nociones tales como "conjunto" están definidas, y describen cómo manipularlas. Para establecer un marco riguroso en el que se discuta la hipótesis del continuo, es necesario especificar un sistema de axiomas para la teoría de conjuntos.

En 1964, Paul Cohen demostró que la veracidad de la hipótesis del continuo depende de cuáles axiomas se escojan para la teoría de conjuntos. Esta situación es similar a la de la veracidad o falsedad del axioma de las paralelas de Euclides, lo cual depende del tipo de geometría: hay una geometría "euclidiana" para la cual es verdadero, pero también hay geometrías "no euclidianas" para las cuales es falso (véase página 252). De manera similar, hay teorías de conjuntos "cantorianas" en las que la hipótesis del continuo es verdadera, y teorías de conjuntos "no cantorianas" en las que es falsa. Kurt Gödel había demostrado que la hipótesis del continuo es verdadera en algunas axiomatizaciones de la teoría de conjuntos, y Cohen, usando una nueva técnica llamada *forcing*, demostró que en otras axiomatizaciones es falsa. En particular, no hay una elección de axiomas distinguida que conduzca a una única teoría de conjuntos "natural".

§5. Notación de teoría de conjuntos

[*véase página* 140]

La notación matemática sigue modas, y a veces la moda puede cambiar. En consecuencia, la terminología de Courant y Robbins en ocasiones difiere en detalles menores de la acostumbrada actualmente, pero esto rara vez es suficientemente importante como para mencionarlo. En este caso particular, sin embargo, la diferencia con el uso actual es demasiado significativa como para ignorarla.

Los términos "suma lógica" y "producto lógico" prácticamente no se usan en nuestros días; en su lugar se emplean las alternativas "unión" e "intersec-

ción". El conjunto vacío se denota como $\varnothing$ y no como O, y ya no se utiliza un símbolo especial, I, para el universo de discurso. Las notaciones actuales para la unión e intersección de dos conjuntos A y B son las siguientes:

Unión: $A \cup B$ (en lugar de la notación de Courant y Robbins, $A + B$)

Intersección: $A \cap B$ (en lugar de la notación de Courant y Robbins, AB).

El complemento A' suele escribirse A^c, pero todavía es común usar A'. La notación actual para subconjuntos es cualquiera de los dos símbolos: $\subset$ o $\subseteq$. A diferencia de $<$ y $\leq$, la expresión $A \subset B$ no implica que $A \neq B$, ni hoy ni en los tiempos de Courant y Robbins. Para denotar desigualdad en una relación de subconjuntos, se utiliza la incómoda notación $A \subsetneq B$.

Las notaciones $A + B$, AB y A' sobreviven aún en las ciencias de la computación y en la ingeniería electrónica, en las que se utilizan para describir circuitos formados a partir de entradas lógicas.

Irónicamente, la notación moderna oscurece las analogías algebraicas de las propiedades 6 a 17 de la página 140. Sin embargo, en vista de las propiedades 10, 11 y 13, puede que esto no sea del todo grave.

§6. El teorema de los cuatro colores
[*véanse páginas* 280 y 299]

El teorema de los cuatro colores fue demostrado en junio de 1976 por Kenneth Appel y Wolfgang Haken. Su demostración depende de que se muestre que unos dos mil mapas específicos se comportan de un modo particular algo complicado. Examinar todos esos casos resulta enormemente tedioso, así que usaron una computadora, la cual requirió de varios miles de horas para completar las verificaciones. Ahora la demostración puede verificarse en pocas horas, gracias a mejores métodos teóricos y a computadoras más rápidas, pero todavía no se ha encontrado alguna demostración hecha a "lápiz y papel". ¿Existe una demostración más sencilla? Nadie sabe, aunque se ha visto que ninguna demostración sustancialmente más sencilla puede darse siguiendo líneas similares.

La demostración de Courant y Robbins del teorema de los cinco colores (página 299) es una adaptación del trabajo de Arthur Kempe, un abogado y matemático aficionado, quien en 1879 publicó una presunta demostración del teorema de los cuatro colores. Utiliza una variante del método de inducción matemática (páginas 32 a 43), tomando como idea básica que si el teorema de los cuatro colores es falso entonces debe haber mapas que requieren un quinto color. Si tales mapas "malos" existen, pueden incorporarse a mapas más grandes en todas las maneras posibles, y todos ellos requerirán también

un quinto color. Como no tiene caso alguno hacer más grandes los mapas malos, hay que ir en sentido opuesto y examinar los mapas malos más pequeños. Coloquialmente a un mapa así se le denomina *minimal criminal.*† La existencia de un *minimal criminal* se sigue del principio del menor entero (página 42), que es equivalente al principio de inducción matemática. Un *minimal criminal* se distingue por las siguientes dos propiedades: necesita cinco colores, y cualquier mapa con un número menor de países necesita sólo cuatro. La demostración procede haciendo uso de estas propiedades para restringir la estructura de un *minimal criminal,* hasta que finalmente se muestra que no existe *minimal criminal* alguno. Por contradicción (demostración indirecta, véase la página 113), el teorema debe ser verdadero.

La idea de Kempe era partir de un *minimal criminal* y producir un mapa relacionado más pequeño, el cual, por la segunda propiedad mencionada anteriormente, puede iluminarse con cuatro colores; de ahí, Kempe trató de deducir que el mapa original también se puede iluminar con cuatro colores: la contradicción requerida. Específicamente, el procedimiento consistía en tomar un *minimal criminal* y contraer a un punto alguna región convenientemente seleccionada; el mapa resultante tiene menos regiones, de manera que puede iluminarse con cuatro colores. Podría darse el caso de que fuera imposible regenerar la región contraída y encontrar un color para ella sin cambiar los colores del resto del mapa, a causa de que dicha región podría lindar con otras que ya usan los cuatro colores. Sin embargo, si la región que se contrae es un triángulo (una región que linda sólo con otras tres), no hay problema. Si es un cuadrado entonces se emplea una técnica artificiosa de intercambio de colores, llamada ahora "cadena de Kempe", para cambiar uno de los colores vecinos, lo cual permite que la estratagema funcione. Si es un pentágono, mantuvo Kempe, funciona un argumento similar. Finalmente, pudo demostrar que todo mapa debe contener ya sea un triángulo, un cuadrado o un pentágono, así que siempre hay una región adecuada para ser contraída y regenerada.

En 1890, Percy Heawood encontró un error en el tratamiento de Kempe de las regiones pentagonales. Heawood notó que el método de Kempe puede remendarse para dar una demostración de que cinco colores son siempre suficientes: un color adicional hace más fácil regenerar el pentágono. Ésta es la demostración presentada en la página 299. Por otra parte, nadie podía encontrar un mapa que realmente necesitara cinco colores.

En 1922, Philip Franklin demostró que todo mapa con 26 o menos regiones se puede iluminar con cuatro colores. Su método sentó las bases para el exitoso asalto final, con la idea de una configuración reductible. Una *configu-*

† Literalmente, "criminal mínimo" o "menor infractor". En lo que sigue, se usará el término empleado en la versión original en inglés, *minimal criminal,* para denominar a tal mapa que, a fin de cuentas, no existe. [N. E.]

ración es sólo un conjunto de regiones conectadas del mapa junto con información de cuántas regiones son adyacentes a cada una alrededor del exterior. Para ver lo que significa reductibilidad, considérese el ejemplo de contraer y regenerar una región triangular. Contráigase el triángulo a un punto y supóngase que el mapa resultante, que tiene una región menos, puede iluminarse con cuatro colores; lo mismo puede hacerse entonces con el mapa original, pues el triángulo linda sólo con tres regiones y eso deja un cuarto color libre cuando dicho triángulo es regenerado en el mapa. De manera más general, una configuración es *reductible* si puede demostrarse que la iluminación con cuatro colores de cualquier mapa que la contiene es posible siempre que un mapa más pequeño pueda iluminarse con cuatro colores. Con un argumento similar se demuestra que los cuadrados son reductibles. Kempe pensó que los pentágonos eran reductibles, pero estaba equivocado.

Evidentemente un *minimal criminal* no puede tener una configuración reductible, así que si mostramos que todo *minimal criminal debe* contener una configuración reductible, tenemos la contradicción requerida. El modo más directo de hacerlo es encontrar un conjunto de configuraciones reductibles que sea *inevitable*, en el sentido de que cualquier mapa (no sólo un *minimal criminal*) debe contener una de tales configuraciones. Kempe había intentado hacer eso precisamente: demostró que el conjunto {triángulo, cuadrado, pentágono} es inevitable, lo cual es correcto, pero cometió un error en la demostración de la reductibilidad del pentágono. Con todo, la estrategia básica de su demostración —encontrar un conjunto inevitable de configuraciones reductibles— fue una idea brillante.

En 1950, Heinrich Heesch se convirtió en el primer matemático en afirmar públicamente que creía que el teorema de los cuatro colores podría demostrarse encontrando un conjunto inevitable de configuraciones reductibles. Se dio cuenta, sin embargo, de que el conjunto inevitable tendría que contener muchas más configuraciones que las tres del intento fallido de Kempe, pues el pentágono ha de ser reemplazado por una copiosa lista de alternativas. De hecho, Heesch estimó que se necesitarían alrededor de 10 000 configuraciones, cada una de ellas de tamaño moderado. Además, inventó un método para demostrar la inevitabilidad, basado en una vaga analogía eléctrica. Supongamos que se aplica una cierta cantidad de carga eléctrica a cada región, y luego se le permite moverse a regiones vecinas siguiendo diversas reglas. Por ejemplo, podemos determinar que la carga sobre cualquier pentágono se divida en partes iguales y se transfiera a cualquiera de sus vecinos, excepto triángulos, cuadrados y pentágonos. A partir del análisis de los rasgos generales de las distribuciones de carga, se puede mostrar que deben darse ciertas configuraciones específicas, pues de otra forma la carga "se esfumaría". Recetas más complicadas llevan a listas más complicadas de configuraciones inevitables.

En 1970 Wolfgang Haken hizo mejoras al método de cargas de Heesch y empezó a pensar seriamente en resolver el problema de los cuatro colores. La mayor dificultad era el tamaño probable de configuraciones en el conjunto inevitable. Si partimos de un estimado de 10 000 regiones para ser examinadas en cuanto a su reductibilidad, el cómputo total podría fácilmente tomar un siglo. Y si, a fin de cuentas, una sola configuración del conjunto inevitable resultara no ser reductible, todo el cómputo habría sido inútil.

Entre 1972 y 1974, Haken y Kenneth Appel comenzaron juntos un diálogo interactivo con la computadora para tratar de mejorar las posibilidades de éxito. La primera versión de su programa de computadora proporcionó mucha información útil. Modificaron después el programa para superar una serie de defectos e intentaron nuevamente. Aparecieron problemas más sutiles y fueron debidamente corregidos. Después de unos seis meses de este diálogo, Appel y Haken se convencieron de que su método para demostrar la inevitabilidad tenía buenas posibilidades de éxito. En 1975, su programa de investigación pasó de la fase de exploración al ataque final. En enero de 1976, empezaron la construcción de un conjunto inevitable con unas 2 000 regiones, y para junio de 1976 su trabajo estaba completo. Entonces comprobaron la reductibilidad de cada configuración de este conjunto. Aquí la computadora demostró ser indispensable, reportando puntualmente que cada una de las 2 000 configuraciones del conjunto inevitable de Appel y Haken es reductible. Esto contradice la supuesta existencia de un *minimal criminal*, de manera que cuatro colores solos bastan para iluminar cualquier mapa plano.

¿Hasta qué punto puede un argumento que descansa en un cómputo enorme —que ningún cerebro humano sin ayuda podría comprobar— ser considerado una demostración? Stephen Tymoczko, un filósofo, escribió: "Si aceptamos el teorema de los cuatro colores como un teorema, nos vemos forzados a cambiar el sentido de 'teorema', o más específicamente, a cambiar el sentido del concepto subyacente de 'demostración' ". Pocos investigadores en matemáticas están de acuerdo con esto. Una razón es que existen demostraciones matemáticas que no están basadas en una computadora y sin embargo son tan largas y complicadas que ni siquiera después de estudiarlas por una década podría alguien meter las manos al fuego y declarar que están totalmente exentas de errores. Por ejemplo, el llamado "teorema de clasificación de grupos simples finitos" tiene al menos 10 000 páginas, requirió los esfuerzos de más de cien personas y sólo un especialista bien entrenado puede seguirlo. Sin embargo, los matemáticos generalmente están convencidos de que la demostración es correcta. La razón es que la estrategia tiene sentido, los detalles cuadran entre sí, nadie ha encontrado un error serio y el juicio de la gente que hace el trabajo es por lo menos tan confiable como el de alguien no versado en el tema. Es evidente que esta convicción se desvane-

cería si cualquiera —conocedor o no del tema— encontrara un error, pero hasta ahora no ha sucedido.

No hay nada en la demostración de Appel y Haken que sea menos convincente que el teorema de clasificación para grupos simples finitos. De hecho, es mucho menos probable que una computadora cometa un error a que lo haga un humano, siempre y cuando el programa sea correcto. La estrategia de demostración de Appel y Haken tiene mucho sentido desde el punto de vista lógico; su conjunto inevitable fue obtenido en todo caso a mano, y parece haber pocas razones para dudar de la exactitud del programa utilizado para comprobar la reductibilidad. Pruebas hechas al azar no han encontrado nada fuera de lugar. En una entrevista para el periódico, Haken resumió el punto de vista general: "cualquiera puede, en cualquier parte del proceso, llenar los detalles y verificarlos. El hecho de que una computadora pueda comprobar más detalles, en pocas horas, de los que un humano podría llegar a esperar comprobar en una vida no cambia el concepto básico de demostración matemática. Lo que ha cambiado no es la teoría sino la práctica de las matemáticas".

§7. La dimensión de Hausdorff y los fractales

[*véase página* 282]

La definición de dimensión que dio Poincaré en 1912 (véase página 284) es topológica y —lo que es muy razonable— conduce siempre a un valor entero. Un concepto de dimensión muy diferente del de Poincaré ha adquirido importancia recientemente. Fue inventado originalmente por Félix Hausdorff en 1919 y desarrollado por A. S. Besicovitch en la década de 1930, pero después fue abandonado en las investigaciones matemáticas. Se ha puesto de moda otra vez por sus aplicaciones a la teoría de Benoît Mandelbrot de los *fractales* —objetos geométricos con estructura en todas las escalas de amplificación, tales como el famoso *conjunto de Mandelbrot* (Figura 288)—.

Este conjunto consta de todos los números complejos c (los cuales pueden representarse como puntos del plano) tales que la sucesión c, c^2+c, $(c^2+c)^2+c$, ... no tienda a infinito; cada término de esta sucesión es el cuadrado del término anterior más c.

La dimensión Hausdorff-Besicovitch de un conjunto, con frecuencia denominada ahora la *dimensión fractal,* tiene muchas aplicaciones en diferentes ramas de la ciencia, ya que es una cantidad precisa que puede medirse experimentalmente y compararse con la teoría. Sorprendentemente, no necesita ser un entero. Este rasgo curioso, la razón por la que todavía tiene sentido considerar a este número como una dimensión, puede entenderse partiendo de una

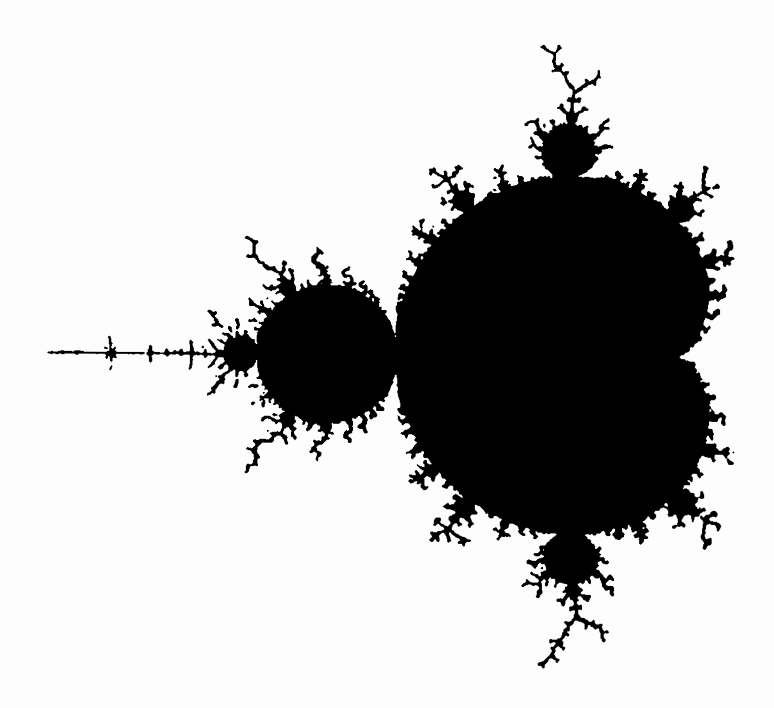


FIGURA 288. *El conjunto de Mandelbrot tiene una estructura intrincada en todas las escalas de amplificación.*

versión más sencilla, conocida como *dimensión de reescalamiento.* Algunas figuras pueden ensamblarse para formar copias más grandes de sí mismas. Por ejemplo (véase Figura 289), se requieren dos copias de un segmento de línea recta (un objeto unidimensional) para construir un segmento de línea recta del doble de tamaño; se requieren cuatro copias de un cuadrado (bidimensional) para hacer uno del doble de tamaño, y se requieren ocho copias de un cubo (tridimensional) para hacer uno del doble de tamaño. En general, se requieren 2^d copias de un hipercubo d-dimensional (véase página 265) para hacer uno del doble de tamaño, y se requieren $c = a^d$ copias para hacer uno de a veces su tamaño.

Podemos resolver la ecuación $c = a^d$ para d empleando logaritmos (véase página 486, ecuación (6)):

$$\log c = d \log a,$$

de manera que

$$d = \frac{\log c}{\log a}. \tag{6}$$

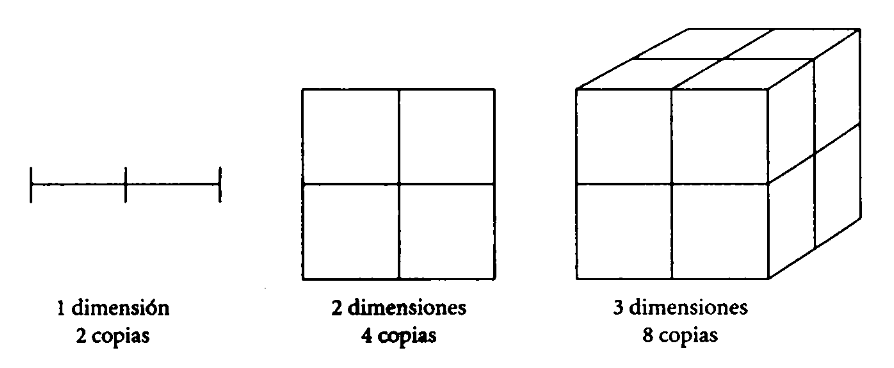


FIGURA 289. *El número de copias requeridas para duplicar el tamaño de un objeto depende de su dimensión.*

Podemos ir ahora en el otro sentido y utilizar esta ecuación para definir d, dados c y a. El resultado se conoce como la dimensión de reescalamiento del conjunto en cuestión. Si estudiamos ejemplos, veremos que esto conduce a conclusiones intrigantes. Tomemos por caso el conjunto de Cantor (véase página 282), que puede hacerse del triple de su tamaño ($a = 3$) ensamblando dos copias ($c = 2$) (véase Figura 290).

De acuerdo con la definición (6), la dimensión de reescalamiento del conjunto de Cantor es entonces

$$d = \frac{\log 2}{\log 3} = 0.630923\ldots,$$

un número real pero que no es un entero. De manera similar, el triángulo de Sierpiński (Figura 291) puede duplicarse en tamaño ($a = 2$) ensamblando tres copias, de manera que su dimensión de reescalamiento es

$$d = \frac{\log 3}{\log 2} = 1.584962\ldots.$$

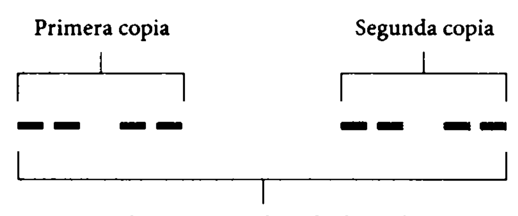


FIGURA 290. *Dos copias del conjunto de Cantor triplican su tamaño.*

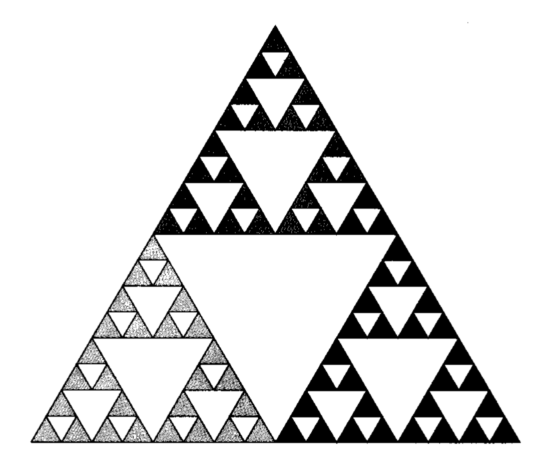


FIGURA 291. *Tres copias de un triángulo de Sierpiński doblan su tamaño.*

Esta cantidad se considera una dimensión porque toma el mismo valor que la dimensión común para conjuntos "bien portados" tales como intervalos, cuadrados, cubos, y así sucesivamente. La dimensión fractal concuerda con la dimensión de reescalamiento para muchos conjuntos, pero está definida para conjuntos que no pueden agrandarse ensamblando copias de sí mismos. Generalmente, la dimensión fractal de un conjunto fractal no es un entero, aunque a veces puede serlo; por ejemplo, en 1991, Mitsuhiro Shishikura demostró que la dimensión fractal de la frontera del conjunto de Mandelbrot es 2. El verdadero significado de la dimensión fractal radica en ser una medida de "qué tan bien es llenado el espacio por el conjunto" o "qué tan áspero es el conjunto". Por ejemplo, el conjunto de Cantor, con una dimensión estrictamente entre 0 y 1, llena el espacio mejor que un punto (dimensión 0) pero no mejor que un segmento de línea recta (dimensión 1). Así, la dimensión fractal resuelve la cuestión de si el conjunto de Cantor debiera tener dimensión 0 o 1 (véase página 283) de una manera muy diferente a la propuesta de Poincaré.

§8. NUDOS

[*véase página* 290]

La teoría de nudos es actualmente el foco de una gran cantidad de actividad en la investigación, impulsada por el descubrimiento del polinomio de Jones, un nuevo método notable para distinguir nudos que no son equivalentes topológicamente. La teoría abarca tanto enlaces como nudos, así que comenzaremos por precisar más estos conceptos.

Un *enlace* es un conjunto de una o más gazas cerradas en el espacio tridimensional. Las gazas individuales se llaman *componentes* del enlace. Las gazas pueden torcerse o anudarse, y —como el nombre sugiere— es posible unirlas entre sí de cualquier manera, incluyendo el caso de no estar unidas en lo absoluto en el sentido convencional. Si hay sólo una gaza, el enlace se llama *nudo*. El problema central en la teoría de enlaces es encontrar modos eficientes para decidir si dos enlaces o nudos dados son o no equivalentes topológicamente —es decir, si pueden deformarse uno en el otro mediante transformaciones continuas (véase páginas 275-276)—. En particular, queremos saber si lo que parece un nudo está en realidad desanudado, es decir, si es equivalente al *no-nudo* (Figura 292a), y si un enlace dado de n componentes puede ser desenlazado, es decir, si es equivalente al *no-enlace* de n componentes (Figura 292b).

La manera de lograr lo anterior es encontrando las denominadas *invariantes topológicas*, que son números —u objetos matemáticos más complicados— que no cambian cuando el enlace es deformado continuamente. Así, enlaces con invariantes distintas no deben ser equivalentes topológicamente; sin embargo, enlaces con las mismas invariantes pueden o no ser equivalentes, y la única forma de decidirlo es encontrando una equivalencia topológica o inventando una invariante más sensible.

La invariante de nudo estándar en la teoría de nudos en la era anterior a Jones era el *polinomio de Alexander*, inventado en 1926. Ésta le asigna a cada nudo un polinomio en una variable t, que puede calcularse siguiendo un procedimiento estándar. No necesitamos ocuparnos aquí del procedimiento preciso, sino indicar el tipo de resultados que se obtienen. La Figura 293 muestra varios nudos simples y sus polinomios de Alexander.

El polinomio de Alexander es suficientemente bueno para distinguir entre un nudo de trébol y un nudo de arrecife, ya que éstos tienen diferentes polinomios de Alexander, pero *no* es suficientemente bueno como para distinguir entre un nudo de arrecife y un nudo "de la abuelita", o entre un trébol a izquierda y un trébol a derecha, aunque sea "obvio" experimentalmente que

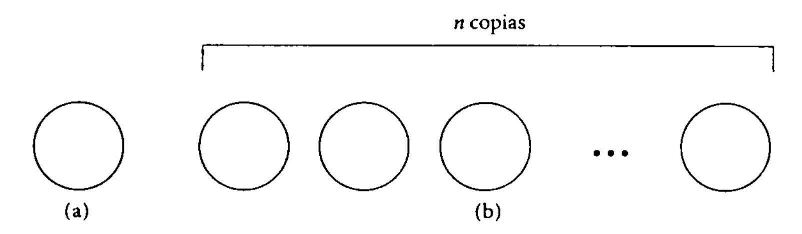
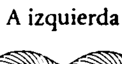
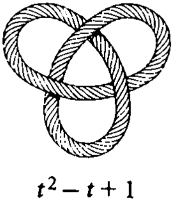


FIGURA 292. *a) el no-nudo; b) el no-enlace de* n *componentes.*

Tréboles

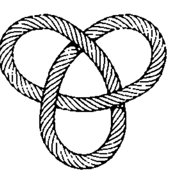
A izquierda





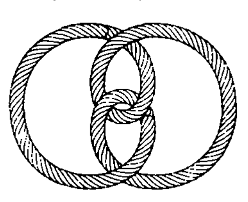
$t^2 - t + 1$

A derecha



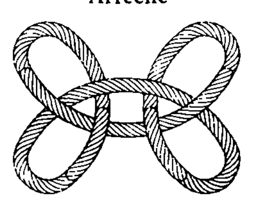
$t^2 - t + 1$

Forma de ocho



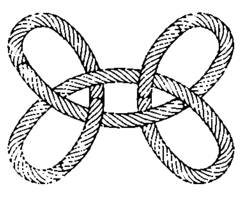
$t^2 - 3t + 1$

Arrecife



$t^4 - 2t^3 + 3t^2 - 2t + 1$

Nudo de la abuelita



$t^4 - 2t^3 + 3t^2 - 2t + 1$

FIGURA 293. *Algunos nudos comunes y sus polinomios de Alexander.*

estos nudos no son equivalentes en realidad. El problema es, ¿cómo podemos demostrar esto? De 1926 a 1984, los matemáticos se esforzaron por resolver estas cuestiones y otras similares. Lo lograron, pero mediante métodos complicados. La teoría de nudos no llegó precisamente a estancarse pero ciertamente necesitaba nuevas ideas.

En 1984 Vaughan Jones, un neozelandés, estaba trabajando sobre cuestiones de análisis con las llamadas funciones de traza en álgebras de operadores, que habían surgido en relación con la física matemática. D. Hatt y Pierre de la Harpe se dieron cuenta de que algunas de las ecuaciones de Jones parecían más bien ecuaciones que tiene lugar en teoría de trenzas (las trenzas son sistemas de líneas unidos, muy relacionados con los enlaces). Ponderado las razones que podrían estar detrás de tal coincidencia, Jones descubrió que sus funciones de traza podían utilizarse para definir una invariante polinomial para enlaces.

En un principio se pensó que el polinomio de Jones debería ser sólo alguna variación del polinomio de Alexander, pero pronto se hizo claro que era genuinamente nuevo. Posteriormente se encontraron definiciones más simples en las que no se recurre a álgebras de operadores. Cinco grupos separados de matemáticos descubrieron de manera simultánea e independiente una generalización que era aún mejor para distinguir nudos, una fórmula de dos variables llamada con frecuencia el polinomio HOMFLY, por los nombres de sus descubridores: *H*oste-*O*cneanu-*M*illet-*F*reyd-*L*ickorish-*Y*etter. Hoy día existen una docena o más de nuevos polinomios de nudos que han resuelto muchos problemas notables, pero a la vez han planteado muchos rompecabezas nuevos por sí mismos, ya que no se adaptan cómodamente al aparato topológico establecido. En cierto sentido, aunque los topólogos pueden calcular los polinomios de nudos y demostrar teoremas acerca de ellos, todavía no están seguros de qué son realmente estas nuevas invariantes polinomiales, aunque parecen tener alguna relación profunda con la física cuántica.

El polinomio original de Jones es una invariante suficientemente poderosa para distinguir un trébol a izquierda de uno a derecha, lo que el polinomio de Alexander no podía determinar. El polinomio HOMFLY es aún más poderoso, y puede distinguir un nudo de arrecife de un nudo de la abuelita. De hecho, si denotamos con $P(L)$ al polinomio HOMFLY de un enlace, tenemos

$$\begin{aligned}
P(\text{trébol a izquierda}) &= -2x^2 - x^4 + x^2y^2, \\
P(\text{trébol a derecha}) &= -2x^{-2} - x^{-4} + x^{-2}y^2, \\
P(\text{nudo de arrecife}) &= (-2x^2 - x^4 + x^2y^2)(-2x^{-2} - x^{-4} + x^{-2}y^2), \text{ y} \\
P(\text{nudo de la abuelita}) &= (-2x^2 - x^4 + x^2y^2)^2.
\end{aligned}$$

En estos polinomios, x y y son las dos variables requeridas para definirlos. Estos resultados obviamente demuestran no sólo que los dos tipos de trébol no son equivalentes topológicamente sino que el nudo de arrecife y el nudo de la abuelita tampoco lo son.

§9. Un problema de mecánica

[*véase página* 357]

Es éste el único caso sobre el que puede argüirse que Courant y Robbins cometieron un error, aunque añadiendo más condiciones es posible salvar su argumento. Paradójicamente, es más fácil detectar el hueco en su demostración si adoptamos un enfoque topológico de la dinámica, que era por lo que intentaba abogar el argumento de ellos.

Repetimos el enunciado del problema. Supóngase que un tren viaja entre dos estaciones siguiendo una vía recta. Con una bisagra se une una varilla al piso de uno de los vagones, de manera que pueda moverse sin fricción ya sea hacia adelante o hacia atrás hasta que toque el piso (Figura 175, página 358). Si toca el piso, supóngase que permanece ahí durante todo el movimiento subsiguiente; supóngase también que de antemano especificamos cómo se mueve el tren. El movimiento no tiene que ser uniforme: el tren puede acelerar, detenerse repentinamente o incluso ir en reversa durante un tiempo determinado y debe empezar en una estación y terminar en la otra.

Courant y Robbins preguntan si es siempre posible colocar la varilla en una posición tal que nunca toque el piso durante el viaje. Su solución consiste en hacer notar que la posición final de la varilla depende continuamente de su posición inicial. Como hay un rango continuo de ángulos iniciales, de 0° a 180°, y la posición final depende continuamente de la posición inicial, el teorema de Bolzano (página 350) implica que el rango de ángulos finales también es continuo. Si empezamos con la varilla caída hacia adelante a 0°, ahí permanecerá; si empezamos con la varilla caída hacia atrás a 180°, ahí se quedará. Así que el rango de ángulos finales incluye todos los valores entre 0° y 180°; en particular incluye a 90°, así que podemos acomodar la varilla para que termine en posición vertical. Como la varilla se quedaría en el piso al tocarlo, debe evitarse que esto ocurra.

La dificultad radica en que puede argüirse que la suposición de continuidad en la discusión anterior no está justificada. El problema no radica en lo sofisticado de las leyes de Newton del movimiento, sino en las "dominantes condiciones en la frontera": si la varilla toca el piso, se queda ahí. Para ver por qué las condiciones en la frontera causan problemas, introducimos una representación topológica de los movimientos posibles del sistema. Este enfoque,

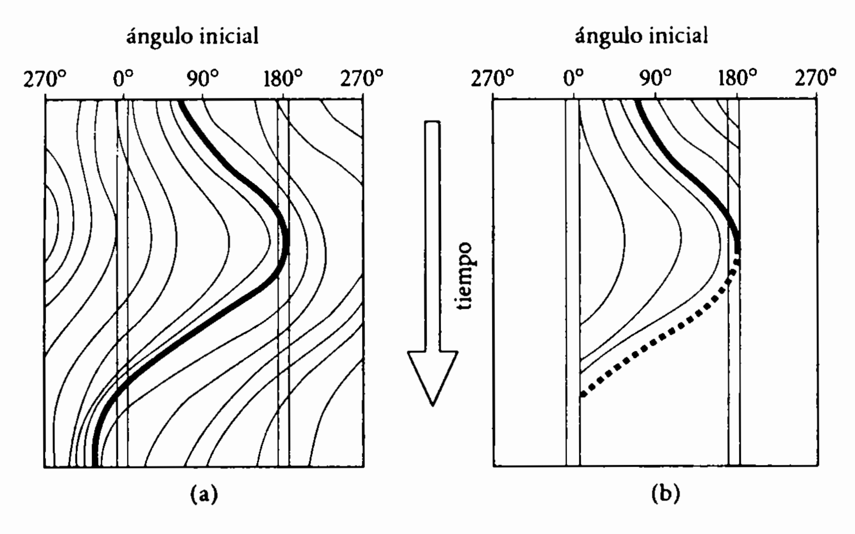


FIGURA 294. *Historias posibles de la varilla móvil para diferentes condiciones iniciales: a) sin condiciones en la frontera; b) lo que sucede cuando se imponen condiciones en la frontera.*

conocido como *retrato de fase*, se remonta a Poincaré. La idea consiste en dibujar un tipo de diagrama espacio-tiempo del movimiento, no sólo para una única posición inicial de la varilla sino para muchas posiciones diferentes —en principio, para todas—. La posición de la varilla es un ángulo entre 0° y 360°, y la podemos graficar en la dirección horizontal (véase Figura 294). Dejemos que el tiempo transcurra en la dirección vertical. Nótese que los bordes izquierdo y derecho de esta figura deben empatarse, ya que 0° = 360°: conceptualmente, el rectángulo se enrolla para hacerlo un cilindro.

Ahora, la trayectoria en el espacio y el tiempo del ángulo que determina la posición de la varilla forma alguna curva que corre a lo largo del cilindro —lo que Albert Einstein llamó una "línea de universo"—. Ángulos iniciales diferentes conducen a curvas diferentes; las leyes de la dinámica muestran que estas curvas varían continuamente conforme el ángulo inicial varía continuamente —siempre que no se impongan las condiciones en la frontera—. Sin esas condiciones la varilla es libre de girar 360°: no hay piso que le impida dar la vuelta completa. En la Figura 294a se muestra una posible evolución, y en ella la posición final sí depende continuamente de la posición inicial.

Sin embargo, si se vuelven a imponer las dominantes condiciones en la frontera (Figura 294b), la posición final *no* tiene que depender continuamente de la posición inicial. Curvas que sólo rocen la frontera izquierda pueden girar de regreso hasta la derecha. Ciertamente, en esta figura particular *todas* las posiciones iniciales terminan en el piso: contrario a lo que sostienen

Courant y Robbins, no hay una elección de posición inicial que impida que la varilla caiga al piso durante el movimiento completo.

Este error en los argumentos de Courant y Robbins fue señalado por primera vez por Tim Poston en 1976, pero todavía no es muy ampliamente conocido. La suposición de continuidad puede resucitarse imponiendo más restricciones al movimiento, como por ejemplo una vía perfectamente plana, ningún brinco en el tren y así por el estilo. Sin embargo, parece más instructivo, como un ejercicio de aplicación de la topología a la dinámica, entender por qué las absorbentes condiciones en la frontera destruyen la continuidad. Esta dificultad es importante en la dinámica topológica avanzada, en la que ha dado lugar al concepto de un "bloque aislante", que es una región tal que ninguna trayectoria dinámica es tangente a su frontera.

§10. El problema de Steiner

[*véase página* 398]

El problema de Steiner (página 393) se refiere a un triángulo *ABC* y nos pide que encontremos un punto *P* tal que minimice la distancia total $PA+PB+PC$. La respuesta, al menos cuando los ángulos del triángulo *ABC* son menores que 120°, es que *P* es el único punto tal que las líneas *PA*, *PB* y *PC* se interceptan en ángulos de 120° una con la otra (páginas 393-394). El problema de Steiner puede generalizarse al problema de la red de caminos, el cual pide la red de líneas (caminos) de menor longitud total que une un conjunto dado de puntos (pueblos) uno al otro (página 398). Este problema generalizado ha dado lugar a una conjetura fascinante, sólo recientemente demostrada.

Supóngase que queremos encontrar una red de líneas que conectarán un conjunto de pueblos. Una manera de hacerlo es utilizar una red llamada generadora, que utiliza sólo líneas rectas que unan pares de pueblos; otra, es usar una red *de Steiner,* en la que se permiten pueblos adicionales, de manera que las líneas que los unen se intercepten en ángulos de 120°. Llamemos a la longitud de la menor red generadora para un conjunto dado de pueblos *longitud generadora* y a la longitud de la menor red de Steiner, *longitud de Steiner.* El problema de encontrar la longitud de Steiner es discutido por Courant y Robbins (página 398) bajo el título "Problema de la red de caminos". Obviamente la longitud de Steiner es menor que o igual a la longitud generadora. ¿Cuánto más pequeña puede hacerse?

Supóngase, por ejemplo, que hay tres pueblos en los vértices de un triángulo equilátero de una unidad de lado. La Figura 295 muestra la menor red de Steiner y la menor red generadora. El nuevo punto introducido en el centro se llama *punto de Steiner:* en general, un punto de Steiner es uno en el que tres

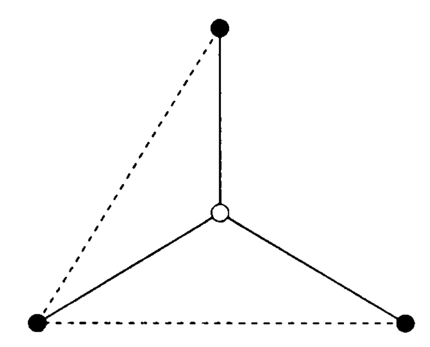


FIGURA 295. *La menor red de Steiner (líneas continuas) y la menor red generadora (líneas interrumpidas) para tres pueblos en un triángulo equilátero.*

líneas (que lo unen con otros puntos en el conjunto de pueblos) se interceptan en ángulos de 120°. La longitud generadora es 2 y la longitud de Steiner es $\sqrt{3}$. En este caso, la razón entre la longitud de Steiner y la longitud generadora es $\sqrt{3}/2 = 0.866$, y el ahorro en distancia obtenido utilizando la menor red de Steiner en vez de la menor red generadora es de aproximadamente 13.34 por ciento.

En 1968, Edgar Gilbert y Henry Pollak conjeturaron que no importa cómo estén localizados los pueblos inicialmente, la longitud de Steiner nunca es menor que la longitud generadora en más de 13.34%; dicho de otra manera:

$$\frac{\text{longitud de Steiner}}{\text{longitud generadora}} \geq \frac{\sqrt{3}}{2} \qquad (7)$$

para cualquier conjunto de pueblos. Esta afirmación se conoce como la *conjetura de la razón de Steiner.* Después de un considerable esfuerzo fue finalmente demostrada por Ding Zhu Du y Frank Hwang en 1991; describiremos su procedimiento una vez que hayamos establecido las bases necesarias.

Encontrar la longitud generadora es un cómputo sencillo, aun para un número muy grande de pueblos. Se resuelve mediante el *algoritmo glotón:* empezamos con la menor línea conectora que podamos hallar, y en cada etapa subsiguiente añadimos la menor línea que encontremos que no complete una figura cerrada; procedemos así hasta que todo pueblo quede incluido. Encontrar la longitud de Steiner es mucho menos fácil, no pueden nada más tomarse todos los ternos posibles de pueblos, encontrar sus puntos de Steiner y buscar la menor red que una los pueblos y esté conectada ya sea en ellos o en dichos puntos de Steiner. Por ejemplo, supóngase que hay seis pueblos acomodados en las esquinas de dos cuadrados adyacentes, como en la Figura 296. Un posi-

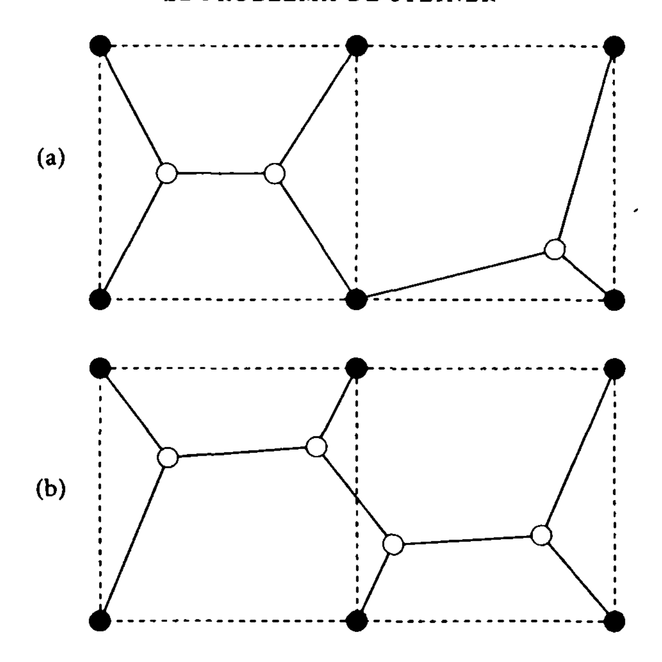


FIGURA 296. *a) Combinación de árboles de Steiner para un cuadrado y un triángulo rectángulo isósceles; b) un árbol de Steiner más corto para el mismo conjunto de pueblos.*

ble árbol de Steiner se muestra en la Figura 296a: se encuentra resolviendo el problema primero para un cuadrado de cuatro pueblos y uniendo después los dos pueblos restantes mediante su punto de Steiner con un pueblo que ya esté conectado. Sin embargo, el menor árbol de Steiner es el que se muestra en la Figura 296b (los cuadrados con puntos negros en los vértices se incluyen sólo para indicar dónde están situados los pueblos).

Por lo visto, no pueden construirse árboles de Steiner mínimos juntando partes. La generalización correcta de la definición de punto de Steiner a un conjunto de muchos pueblos es: cualquier punto en el que los enlaces puedan interceptarse a 120°. Para un ejemplo tan simple como el de cuatro pueblos en los vértices de un cuadrado, dichos puntos no son aquellos puntos de Steiner que conecten algún subconjunto de tres pueblos (Figura 297). Hay una infinidad de puntos en el plano, y aunque la mayoría de ellos probablemente sean irrelevantes, no es obvio que exista algún algoritmo; sin embargo, existen varios: el primero de ellos fue inventado por Z. A. Melzak, pero en la práctica su método es pesado aun para números moderados de pueblos, desde su invención ha sido mejorado, aunque no de manera dramática.

Sabemos ahora que hay razones claras por las que estos algoritmos son ineficientes. Gracias al uso creciente de las computadoras se ha creado una

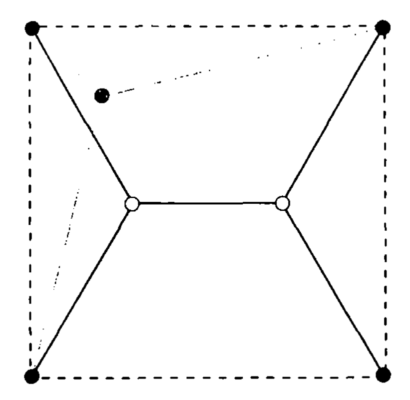


FIGURA 297. *Los puntos de Steiner (en blanco) para cuatro pueblos en un cuadrado (en negro) son diferentes de los puntos de Steiner de un subconjunto de tres pueblos (en gris).*

nueva rama de las matemáticas, la teoría de la complejidad algorítmica. Esta no sólo estudia algoritmos —métodos para resolver problemas— sino también la eficiencia de esos algoritmos. Dado un problema en el que interviene un cierto número n de objetos (aquí, pueblos), ¿cuán rápido crece el tiempo de ejecución de un programa para hallar la solución conforme crece n? Si el tiempo de ejecución no crece más rápido que un múltiplo constante de una potencia fija de n, tal como $5n^2$ o $1066n^4$, entonces se dice que el algoritmo corre en *tiempo polinomial* y se considera que el problema es "fácil"; en general, esto significa que el algoritmo es factible (aunque no lo será si la constante es verdaderamente enorme). Si el tiempo de ejecución crece de manera no polinomial —más rápido que cualquier múltiplo constante de potencias de n, por ejemplo exponencialmente, como 2^n o 10^n— entonces el problema tiene tiempo de ejecución no polinomial y se considera "difícil". En general esto quiere decir que el algoritmo es totalmente impráctico. Entre el tiempo polinomial y el tiempo exponencial hay una maleza de problemas "medianamente fáciles" o "moderadamente difíciles" donde la factibilidad es más un asunto de experiencia.

Por ejemplo, sumar dos números de n dígitos requiere a lo más $2n$ sumas de números de un dígito, incluyendo "lo que se lleva", de manera que el tiempo que toma la solución está acotado por un múltiplo constante (a saber, 2) de la primera potencia de n; la multiplicación larga de dos números n requiere alrededor de n^2 multiplicaciones de números de un dígito y no más de $2n^2$ sumas, o $3n^2$ operaciones con dígitos, de manera que la cota llega ahora sólo a la segunda potencia de n. Estos problemas son, por lo tanto, "fáciles" (aunque un estudiante de primaria discreparía seguramente). En contraste con los ejemplos anteriores, considérese el problema del comerciante ambu-

lante: encontrar la ruta mínima que lleva a un comerciante por un conjunto dado de ciudades. Si hay n ciudades entonces el número de rutas que tenemos que considerar es $n! = n(n-1)(n-2) \ldots 3 \cdot 2 \cdot 1$, que crece más rápido que cualquier potencia de n. Así que la numeración caso por caso es ineficiente más allá de toda esperanza.

Irónicamente, el gran problema de la teoría de la complejidad algorítmica es demostrar que en efecto el tema existe, es decir, demostrar que algún problema "interesante" es en efecto difícil. ¡El problema radica en que es fácil demostrar que un problema es fácil pero difícil demostrar que es difícil! Para mostrar que un problema es fácil, sólo hay que exhibir un algoritmo que lo resuelva en tiempo polinomial, no tiene que ser el mejor ni el más ingenioso: cualquiera servirá. Sin embargo, para demostrar que un problema es difícil no basta con exhibir algún algoritmo cuyo tiempo de ejecución sea no polinomial, ya que puede haberse escogido el algoritmo equivocado y que haya uno mejor cuyo tiempo de ejecución sí sea polinomial. Para eliminar esa posibilidad, hay que encontrar algún modo matemático de considerar todos los algoritmos posibles para el problema y mostrar que ninguno de ellos tiene tiempo de ejecución polinomial, lo cual es extremadamente difícil.

Hay muchos candidatos a problemas difíciles: el problema del comerciante ambulante, el problema de la alacena (¿cómo puede acomodarse mejor un conjunto de objetos de dimensiones dadas en un conjunto de anaqueles de dimensiones dadas?) y el problema de la mochila (dada una mochila de dimensiones fijas y muchos objetos, ¿hay algún conjunto de objetos que llene exactamente la mochila?). Hasta ahora nadie ha conseguido demostrar que alguno de ellos sea difícil. Sin embargo, Stephen Cook, de la Universidad de Toronto, mostró en 1971 que si se puede demostrar que cualquier problema de este grupo de candidatos es en efecto difícil, entonces todos lo son; es decir, se puede "cifrar" cualquiera de ellos para que sea un caso especial de uno de los otros: o todos tienen éxito o todos fallan. Estos problemas se llaman *NP-completos,* donde NP quiere decir no polinomial. Todo mundo *cree* que los problemas NP-completos realmente son difíciles, pero esto nunca ha sido demostrado.

La completividad polinomial se relaciona con el problema de Steiner gracias a que Ronald Graham, Michael Garey y David Johnson han demostrado que el problema de computar la longitud de Steiner es NP-completo. Es decir, cualquier algoritmo eficiente para encontrar la longitud de Steiner precisa para cualquier conjunto de pueblos conduciría automáticamente a soluciones eficientes de todo tipo de problemas de cómputo de los que generalmente se cree que no poseen tales soluciones.

La conjetura de la razón de Steiner (7) es entonces importante, pues demuestra que podemos reemplazar un problema difícil por uno fácil sin perder mucho. Gilbert y Pollak tenían demasiada evidencia positiva cuando enun-

ciaron esta conjetura. Específicamente, podían demostrar una afirmación un poco más débil: la razón entre la longitud de Steiner y la longitud generadora es siempre al menos 0.5. Para 1990 varias personas habían llevado a cabo cálculos heroicos para verificar por completo la conjetura para redes de 4, 5 y 6 pueblos; para arreglos generales de tantos pueblos como se quiera, elevaron los límites de dicha razón de 0.5 a 0.57, 0.74 y 0.8. Alrededor de 1990 Graham y Fang Chung la elevaron a 0.824, mediante un cómputo que describieron como "realmente horrible —estaba claro que no se trataba del enfoque adecuado—".

Para hacer posible el progreso ulterior, tenían que simplificarse los cálculos "horribles". Du y Hwang encontraron una vía mucho mejor: eliminar por completo los cálculos engorrosos. La cuestión básica es cómo hacer que aparezcan triángulos equiláteros en escena. Hay un gran salto desde el ejemplo del triángulo en la Figura 295, que marca la cota de la razón entre la longitud de Steiner y la longitud generadora, a un sistema general de pueblos, que se supone que tiene que respetar la misma cota. ¿Cómo remontar ese vacío? Hay una especie de punto de apoyo intermedio. Imaginemos el plano teselado con triángulos equiláteros idénticos, formando una celosía triangular (Figura 298). Colóquense pueblos sólo en los vértices de las piezas de la teselación. Resulta que los únicos puntos de Steiner que necesitan ser considerados son los centros de las piezas. En pocas palabras, se tiene mucho control, no sólo en los cómputos sino en los análisis teóricos.

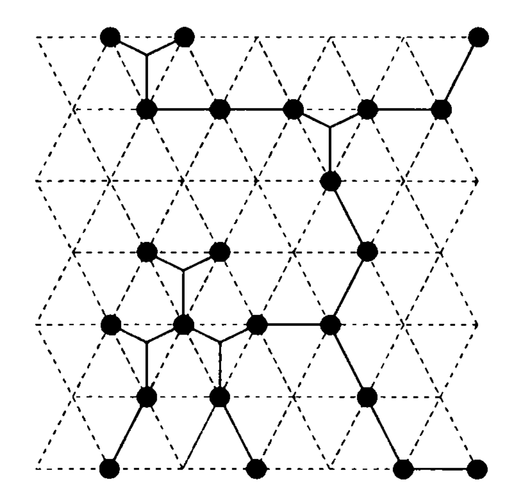


FIGURA 298. *Una red de Steiner para pueblos que están en los vértices de una celosía triangular tiene una estructura mucho más rígida y regular que el caso general. Du y Hwang reducen la conjetura de la razón de Steiner al mismo problema para redes sobre celosías.*

Por supuesto que no todo conjunto de pueblos se acomoda de manera conveniente a una celosía triangular; la creatividad de Du y Hwang consistió en darse cuenta de que los cruciales sí lo hacen. Otra vez la demostración es indirecta, por contradicción. Supóngase que la conjetura es falsa, entonces debe existir un contraejemplo: algún conjunto de pueblos para el cual la razón sea menor que $\sqrt{3}/2$. Du y Hwang muestran que si existe un contraejemplo a la conjetura debe ser uno para el que todos los pueblos queden en una celosía triangular, lo que introduce un elemento de regularidad en el problema, y a partir de ahí es relativamente sencillo completar la demostración.

Para demostrar esta propiedad de las celosías, Du y Wang reformulan la conjetura como un problema en la teoría de juegos, donde los jugadores compiten tratando de limitar las ganancias de sus oponentes. La teoría de juegos fue inventada por John von Neumann y Oskar Morgenstern en su obra clásica, *Theory of Games and Economic Behavior,* de 1947. En la versión de Du y Hwang de la conjetura de la razón de Steiner un jugador selecciona la "forma" general del árbol de Steiner y otro escoge el árbol más corto que pueda encontrarse con esa forma. Du y Hwang deducen la existencia de una celosía que sirve de contraejemplo observando que la ganancia de su juego tiene una propiedad especial de "convexidad".

§11. Películas de jabón y superficies mínimas

[*véase página* 425]

En el capítulo VII, §11, se menciona varias veces la observación de que cuando tres películas de jabón se interceptan parecen formar ángulos de 120°, relacionando este fenómeno con el problema de Steiner (página 393). Se da un fenómeno similar cuando cuatro superficies de película de jabón se interceptan en un punto común, como sucede en la Figura 240, página 426: experimentalmente, el ángulo formado en la esquina de cada superficie es cercano a 109°, siendo éste el valor del ángulo formado por cuatro planos que se interceptan en el centroide de un tetraedro, como en la Figura 299; de paso, esto implica que el pequeño "cuadrado" central de la Figura 240 no es de hecho un cuadrado, lo que explica por qué las trece superficies en el armazón cúbico están ligeramente curvadas.

Estas reglas generales acerca de ángulos fueron registradas por primera vez por Plateau, quien estableció tres principios acerca de la forma que adquieren las películas de jabón en armazones:

1) Constan de un número finito de superficies planas o suavemente curvadas, unidas entre sí de manera fluida.

2) Esas superficies se interceptan sólo de dos maneras: o exactamente tres

de ellas se interceptan a lo largo de una curva suave, o cuatro se interceptan en un punto.

3) Cuando tres superficies se interceptan, los ángulos entre ellas son de 120°; y cuando se interceptan cuatro, los ángulos que se forman en las esquinas son aproximadamente de 109°.

En 1976, Frederick Almgren y Jean Taylor demostraron que estas tres propiedades surgen todas de un mismo principio matemático, sobre el que se fundamenta lo expuesto en §11 del capítulo VII: la película de jabón toma cualquier forma que minimice el área total. Tal vez sorprenda saber que, de los principios de Plateau, el más difícil de establecer es el primero, de carácter más cualitativo: que la figura consta de un número finito de superficies. Los otros dos principios se siguen de una manera relativamente fácil de argumentos geométricos, exactamente igual que el ángulo de 120° en el problema de Steiner. Primero explicamos esta deducción y después discutimos la demostración del primer principio de Plateau.

El paso inicial en la deducción de los principios segundo y tercero a partir del primero consiste en utilizar el hecho de que las superficies son casi planas para reducir el problema a uno acerca de planos. Si se aumenta una región muy pequeña cerca de una línea de intersección de tres superficies o de un punto de intersección de cuatro, las superficies aparecen casi planas, y entre mayor sea la amplificación, más planas parecen ser. Considerando el grado de error que implica tal aproximación, resulta ser suficiente demostrar los principios segundo y tercero de Plateau bajo la suposición simplificadora de que las superficies son planas. El segundo paso es reducir esta cuestión a una referente a líneas sobre una esfera. Considérese la forma en que las regiones planas se intersecan con una esfera con centro en la línea o punto de intersección. El sistema de planos es entonces reemplazado por un sistema de arcos de

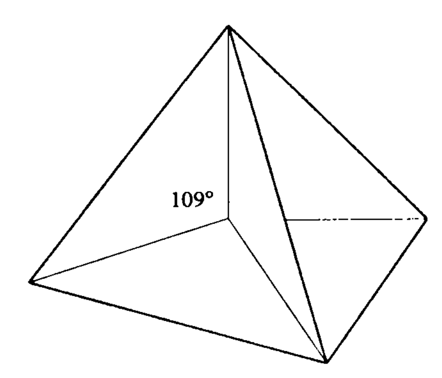


FIGURA 299. *Superficie mínima en un armazón con forma de tetraedro: cuatro superficies se interceptan en el punto central, formando ángulos de 109°.*

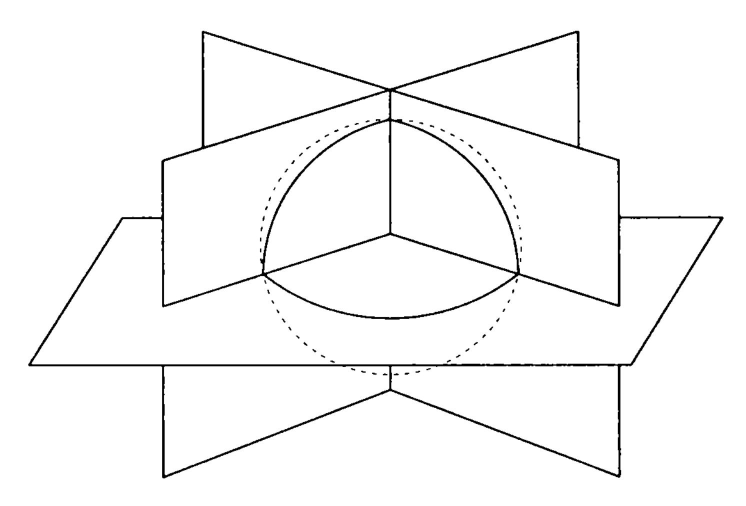


FIGURA 300. *La reducción de la geometría de un sistema de planos a la de un sistema de arcos.*

círculos máximos (véase Figura 300). El requerimiento análogo al de área mínima es que la longitud total de estos arcos debe ser mínima. Por una versión esférica del teorema de Steiner (página 393), demostrada de manera similar, los arcos se interceptan de tres en tres, en ángulos de 120°. El tercer paso es demostrar que precisamente diez configuraciones diferentes de arcos de círculos máximos satisfacen estas condiciones (Figura 301). El cuarto paso es tomar cada una de tales configuraciones y realizar una búsqueda de pequeñas deformaciones de la configuración correspondiente de superficies planas —posiblemente introduciendo nuevas piezas— que reduzcan el área total dentro de la esfera. Si cualquiera de esas reducciones de área es posible, la configuración correspondiente de arcos puede eliminarse: no corresponde a un arreglo de superficies de área mínima. (En la práctica, varios de estos casos fueron estudiados haciendo los armazones de alambre correspondientes y observando las formas que tomaba la película de jabón para deducir la forma general de la pequeña deformación que ocurría. Entonces la posibilidad de reducir el área se establecía de manera rigurosa haciendo estimaciones adecuadas.) Exactamente tres configuraciones sobrevivieron a este proceso: un solo círculo máximo, tres semicírculos que se interceptan en ángulos de 120° y cuatro arcos que forman un tetraedro curvilíneo —números 1 a 3 en la Figura 301—. Las configuraciones planas correspondientes son una sola superficie que no intercepta a ninguna otra, tres superficies que se interceptan en ángulos de 120° o cuatro superficies que se interceptan en ángulos de 109°. Los principios segundo y tercero de Plateau son inmediatos.

Todo depende así de demostrar que la forma mínima consta de un número

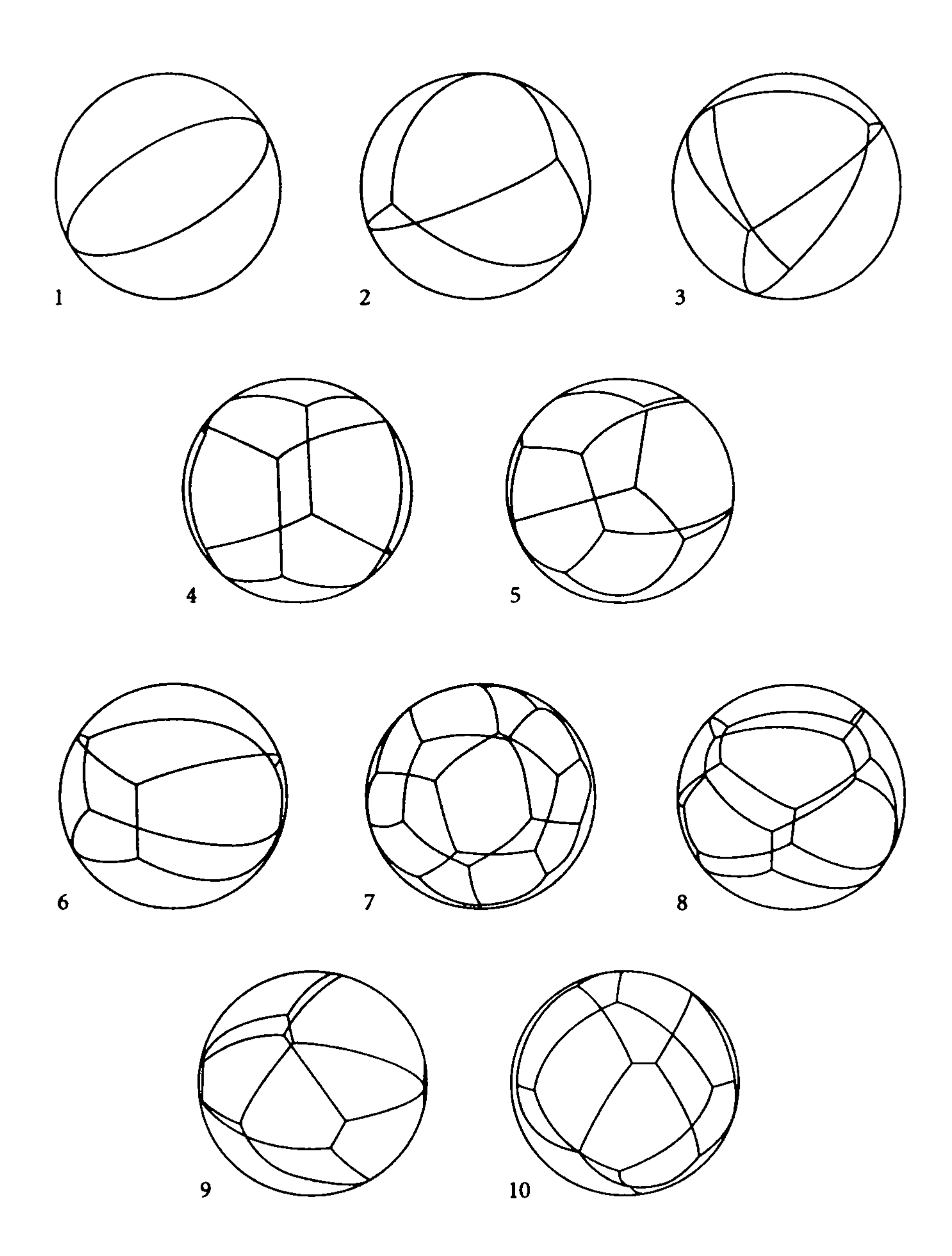


FIGURA 301. *Las diez configuraciones de arcos que se encuentran en ángulos de 120°.*
FUENTE: Scientific American *235, núm. 1 (julio de 1976), pp. 90-91.*

finito de muchas superficies, para lograrlo, es necesario pensar en la posibilidad de que existan formas más complejas, y esto a su vez requiere la generalización del concepto de área para estas formas más complejas. El problema entonces se divide en dos etapas separadas: primera, demostrar la existencia de alguna forma compleja que minimice esta área generalizada, y segunda, usar la propiedad de minimalidad para mostrar que la forma compleja en realidad es bastante simple, y que está compuesta por un número finito de muchas superficies suaves.

Las técnicas para hacer que estas dos etapas funcionen son nuevas y abstractas; pertenecen a un área conocida como "teoría geométrica de la medida" —la misma área a la que pertenece la definición de dimensión fractal—. A grandes rasgos, cualquier superficie particular *S* es reemplazada por una "medida" asociada, una función que asigna a cualquier región *X* del espacio el área de aquella parte de *S* que queda dentro de *X*. Formas más complejas son representadas por funciones con propiedades similares a estas medidas basadas en superficies. La ventaja de reemplazar formas por medidas es que las medidas tienen propiedades mucho más agradables: por ejemplo, pueden sumarse o definirse como el límite de sucesiones de otras medidas, operaciones difíciles de definir directamente para formas geométricas.

La existencia de una medida que minimiza resulta ser entonces un argumento directo en la teoría geométrica de la medida. La parte más difícil del argumento es mostrar que toda medida que minimiza corresponde a un sistema finito de superficies suaves. Irónicamente, el conocimiento acerca de cómo se unirían estas superficies si realmente fueran tales —los principios segundo y tercero de Plateau— ayudaron a Almgren y Taylor a idear cómo demostrar que en efecto son superficies. Sabiendo de antemano cuál "debería ser" la respuesta, con frecuencia hace más fácil encontrar una demostración.

§12. Análisis no estándar

[*véase página* 474]

En la página 476 Courant y Robbins subrayan que "las 'diferenciales' como cantidades infinitamente pequeñas están ahora descartadas definitiva y deshonrosamente": una reflexión precisa del punto de vista que se tenía por consenso cuando se escribió *¿Qué son las matemáticas?* A pesar del veredicto de Courant y Robbins, siempre ha habido algo intuitivo y llamativo en los argumentos a la antigua con infinitesimales. Están aún sumergidos en nuestro lenguaje en ideas tales como "instantes" de tiempo, velocidades "instantáneas" y el considerar a una curva como una serie de líneas rectas infinitamente pequeñas y al área acotada por una curva como suma de una cantidad infinita

de áreas de rectángulos infinitesimales. Este tipo de intuición resulta estar justificado, pues se ha descubierto recientemente que el concepto de cantidades infinitamente pequeñas no es deshonroso y no tiene por qué ser descartado. Es posible establecer un marco riguroso para el análisis en el que las definiciones weierstrassianas en términos de épsilon y delta (véase página 341) sean reemplazadas por enunciados sobre infinitesimales, que son increíblemente similares a las ideas intuitivas de Leibniz, Newton y Cauchy.

La manera de volver respetables a los infinitesimales se llama análisis no estándar y es perfectamente viable como una alternativa al enfoque en términos de épsilon y delta, pero por varias razones —sólo una de las cuales es el conservadurismo científico— la mayoría de los matemáticos aún prefieren el punto de vista de Weierstrass. El gran problema psicológico es que para establecer tal marco se requieren ideas sofisticadas de la lógica matemática moderna. Más o menos de 1920 a 1950 hubo un desarrollo explosivo de la lógica matemática en el que uno de los temas que surgieron fue la *teoría de modelos,* que construye y caracteriza *modelos* de sistemas de axiomas: estructuras matemáticas que obedecen esos axiomas; así, el plano coordenado es un modelo para los axiomas de la geometría euclidiana, el disco de Poincaré (página 258) es un modelo para los axiomas de la geometría hiperbólica, y así sucesivamente.

Hay un sistema de axiomas estándar para los números reales, y se sabe desde hace mucho que hay un modelo único, los números reales estándares R, razón por la cual diferentes maneras de construir "los" números reales (véanse páginas 94-95) conducen a sistemas de números que son efectivamente idénticos. Además, R no contiene ningún infinitesimal ni ningún infinito. Así que, ¿cómo es posible aplicar la teoría de modelos para construir un sistema de números reales "no estándar" que sí contenga estos objetos extraños? Los lógicos distinguen entre sistemas axiomáticos de "primer orden" y de "segundo orden". En una teoría de primer orden los axiomas expresan propiedades que se requiere que cumplan todos los objetos en el sistema, pero no todos los *conjuntos* de objetos. En una teoría de segundo orden no hay tal restricción. En la aritmética ordinaria, un enunciado tal como

$$x + y = y + x \quad \text{para todo } x \text{ y } y \tag{8}$$

es de primer orden, y así lo son todas las leyes comunes del álgebra; sin embargo, el "axioma arquimediano"

$$\text{si } |x| < \frac{1}{n} \text{ para todos los números naturales } n \text{ entonces } x = 0 \tag{9}$$

es de segundo orden. La mayoría de los axiomas comunes para los números reales son de primer orden, pero la lista incluye algunos que son de segundo

orden. De hecho, el axioma de segundo orden (9) es el crucial que excluye de R tanto a los infinitesimales como a los infinitos. No obstante, resulta que si se debilitan los axiomas de manera que consten sólo de las propiedades de primer orden de R entonces existen otros modelos, incluyendo algunos que violan la propiedad (9) anterior. Sea R^* un modelo tal y llamémoslo el sistema de los números hiperreales. Esta idea, la base del análisis no estándar, fue descubierta por Abraham Robinson alrededor de 1960. Ya hemos visto que hay geometrías no euclidianas y teorías de conjuntos no cantorianas, ahora encontramos que hay sistemas de números no arquimedianos.

El conjunto R^* contiene varios subconjuntos importantes: hay un conjunto de números naturales "estándares", $N = \{1, 2, 3, \ldots\}$, y hay también un sistema más amplio de números naturales "no estándares" N^*; están los enteros estándares Z y una extensión correspondiente a los enteros no estándares Z^*; están los racionales estándares Q y una extensión correspondiente a los racionales no estándares Q^*; hay reales estándares R y reales no estándares (o hiperreales) R^*.

Toda propiedad de primer orden de R tiene una única extensión natural a R^*. Sin embargo, (9) expresa una propiedad de segundo orden, y ésta es falsa en R^*. Los hiperreales contienen infinitos e infinitesimales verdaderos. Por ejemplo, $x \in R^*$ es infinitesimal si y sólo si $x \neq 0$ y $|x| < 1/n$ para todo $n \in N$.† El argumento usual de que "los infinitesimales no existen" demuestra en realidad que los infinitesimales *reales* no existen, es decir, que los infinitesimales en R^* no pertenecen a R. Pero eso es completamente razonable, ya que R^* es más grande que R. Dicho sea de paso, el análogo "correcto" de (9) en R^* es

$$\text{si } |x| < \frac{1}{n} \text{ para toda } n \in N^* \text{ entonces } x = 0, \tag{10}$$

y éste es verdadero. Así que cambiar (9) para referirnos a los números naturales no estándares en vez de a los estándares hace una gran diferencia.

La extensión de reales a hiperreales es sólo un ejemplo más del antiguo juego de extender el sistema de números para que se cumpla una propiedad deseada (véanse páginas 77 a 118). Por ejemplo, los números racionales fueron extendidos a los reales para permitir que 2 tenga una raíz cuadrada, y los números reales fueron extendidos a los números complejos para permitir que -1 tenga una raíz cuadrada. Entonces, ¿por qué no extender los números reales a los números hiperreales para permitir que existan los infinitesimales?

Podemos usar R^* para demostrar teoremas acerca de R, ya que los sistemas de números R y R^* son indistinguibles en lo que se refiere a sus propiedades de

† Vale la pena hacer notar que el autor de esta sección presenta un modelo R^* de números hiperreales en el que el cero no es un infinitesimal. [N. T.; N. E.]

primer orden. Sin embargo, R^* tiene todo tipo de rasgos distintivos nuevos, como la presencia de infinitesimales e infinitos, que pueden explotarse de maneras novedosas. Estos nuevos rasgos distintivos son propiedades de segundo orden, y es por eso que los nuevos sistemas pueden tenerlas aunque los viejos no. Observaciones similares se aplican a los subsistemas N y N^*, Z y Z^* y Q y Q^*.

Unas pocas definiciones darán el sabor de este enfoque. Un número hiperreal es *finito* si [su valor absoluto] es menor que algún real estándar; es *infinitesimal* si [su valor absoluto] es menor que todos los reales estándares positivos. Algo que no sea finito es *infinito*, y algo que no esté en R es *no estándar.* Si x es infinitesimal entonces $1/x$ es infinito, y viceversa.

Nada de esto sería de gran importancia si todo lo que pudiera hacerse fuese inventar un nuevo sistema de números. Sin embargo, aunque R y R^* son diferentes, están íntimamente relacionados. De hecho, todo hiperreal finito x tiene una *parte estándar* única std(x) que está infinitamente cercana a x, es decir, $x - \text{std}(x)$ es infinitesimal. En otras palabras, cada hiperreal finito tiene una expresión única como "real estándar más infinitesimal". Es como si cada real estándar estuviera rodeado por una nube de hiperreales infinitamente cercanos, con frecuencia llamada su *halo.* Cada uno de tales halos rodea a un único real, que por alguna oscura razón se suele llamar su *sombra,* aunque una palabra como "núcleo" o "centro" transmitiría mejor la imagen de lo que se quiere decir. Usando la parte estándar podemos transferir propiedades de R^* a R o viceversa.

Para ver cómo difieren las demostraciones en el análisis no estándar de sus contrapartes estándares, considérese el cálculo de Leibniz de la derivada de la función $y = f(x) = x^2$. Lo que él hace es tomar un número pequeño Δx y formar la razón $[f(x + \Delta x) - f(x)]/\Delta x$. (El acercamiento de Newton fue básicamente el mismo, excepto que usó el símbolo o en lugar de Δx). Siguiendo a Leibniz, calculemos:

$$\begin{aligned}\frac{f(x+\Delta x)-f(x)}{\Delta x} &= \frac{(x+\Delta x)^2-x^2}{\Delta x}\\ &= \frac{x^2+2x\Delta x+(\Delta x)^2-x^2}{\Delta x}\\ &= \frac{2x\Delta x+(\Delta x)^2}{\Delta x}\\ &= 2x+\Delta x.\end{aligned}$$

Leibniz arguyó que como Δx es infinitesimal, puede ignorarse, dejando sólo $2x$. Sin embargo, Δx debe ser distinto de cero para que $[f(x+\Delta x)-f(x)]/\Delta x$ tenga sentido, y en tal caso $2x+\Delta x$ *no* es igual a $2x$. Fue esta dificultad la que llevó al obispo George Berkeley a escribir su famosa crítica *The Analyst, Or a Discourse Addressed to an Infidel Mathematician* [1734], en la que señaló algunas inconsistencias lógicas en los fundamentos del cálculo.

Weierstrass superó las objeciones de Berkeley añadiendo un paso final: tomar el *límite* cuando Δx tiende a cero. (Tanto Leibniz como Newton habían expresado ideas similares, pero no con la misma claridad cristalina que en términos de las ε y δ de Weierstrass.) Como valores de Δx distintos de cero pueden tender a cero, podemos suponer que todos los valores de Δx que aparecen durante el cálculo son distintos de cero, de manera que dividir entre Δx tenga sentido. Entonces tomamos el límite cuando $\Delta x \to 0$ para deshacernos del embarazoso terminó adicional Δx y dejar la respuesta requerida $2x$.

En el análisis no estándar hay una manera más sencilla: tómese x finito y estándar (es decir, sea $x \in R$) y supóngase que Δx es un infinitesimal genuino. En lugar de $2x+\Delta x$ tómese su parte estándar $\text{std}(2x+\Delta x)$, que es $2x$. En otras palabras, defínase la derivada de $f(x)$ como

$$\text{std}\left\{\frac{f(x+\Delta x)-f(x)}{\Delta x}\right\},$$

donde x es un real estándar y Δx cualquier infinitesimal.† La idea aparentemente inocente de tomar la parte estándar es exactamente lo que se necesita para hacer que la derivada sea una función real de x en vez de una función hiperreal de x y Δx, lo cual es una manera perfectamente rigurosa de quitar el término Δx, pues $\text{std}(x)$ es un real definido de manera única. En lugar de barrer abajo del tapete al Δx adicional con tantos argumentos especiosos, es elegantemente borrado.

Un curso de análisis no estándar parece un prolongado alarde de exactamente aquellos errores en los que Courant y Robbins gastaron tantas páginas enseñándonos a evitar. Por ejemplo:

1. Una sucesión s_n converge a un límite L si $s_\omega - L$ es infinitesimal para todo infinito ω. (Compárese con la página 327.)

2. Una función f es continua en x si $f(x+\varepsilon)$ está infinitamente cercano a $f(x)$ (es decir, $f(x+\varepsilon)-f(x)$ es infinitesimal) para todo infinitesimal ε. (Compárese con la página 347.)

3. La función f tiene derivada d en x si y sólo si $[f(x+\Delta x)-f(x)]/\Delta x$ está infinitamente cercano a d para todos los infinitesimales Δx. (Compárese con la página 456.)

† Véase la nota de la página 565.

4. El área de una región limitada por una curva es una suma de una infinidad de rectángulos infinitesimales. (Compárese con la página 443.)

Sin embargo, en el marco del análisis no estándar estos enunciados pueden dotarse de un sentido riguroso.

De hecho, el análisis no estándar no conduce a ninguna conclusión acerca de *R* que difiera del análisis estándar. Es fácil concluir de ahí que no tiene sentido usar el enfoque no estándar, ya que "no conduce a nada nuevo". Pero esta crítica no es concluyente, la pregunta no es: "¿da los mismos resultados?" sino más bien: "¿es una manera más sencilla o más natural de obtener esos resultados?" Como Newton mostró en su *Principia*, todo lo que puede demostrarse con cálculo puede también demostrarse con geometría clásica, lo cual no implica de manera alguna que el cálculo no sirva, y lo mismo vale para el análisis no estándar.

La experiencia sugiere que las demostraciones por medio del análisis no estándar son generalmente más cortas y más directas que las demostraciones clásicas en términos de épsilon y delta, lo cual se debe a que evitan estimaciones complicadas de los tamaños de las cosas, estimaciones que constituyen el grueso de una demostración clásica. El obstáculo principal para la adopción generalizada del análisis no estándar es que para saber apreciarlo se requiere una formación que incluya conocimientos de lógica matemática —muy diferente del análisis tradicional—.

BIBLIOGRAFÍA

REFERENCIAS GENERALES

W. AHRENS: *Mathematische Unterhaltungen und Spiele*, 2.ª edición, 2 vols. Leipzig: Teubner, 1910.

W. W. ROUSE BALL: *Mathematical Recreations and Essays*, 11.ª edición, revisada por H. S. M. Coxeter. Nueva York: Macmillan, 1939.

E. T. BELL: *Historia de las matemáticas*. México. Fondo de Cultura Económica, 1949.

—*Los grandes matemáticos. Desde Zenón a Poincaré. Su vida y sus obras*. Buenos Aires. Losada, 1948.

T. DANTZIG: *Aspects of Science*. Nueva York: Macmillan, 1937.

A. DRESDEN: *An Invitation to Mathematics*. Nueva York: Holt, 1936.

F. ENRIQUES: *Questioni riguardanti le matematiche elementari*, 3.ª edición, 2 vols. Bolonia: Zanichelli, 1924 y 1926.

E. KASNER y J. NEWMAN: *Mathematics and the Imagination*. Nueva York: Simon and Schuster, 1940.

F. KLEIN: *Elementary Mathematics from an Advanced Standpoint*, traducida por E. R. Hedrick y C. A. Noble, 2 vols. Nueva York: Macmillan, 1932 y 1939.

M. KRAITCHIK: *La Mathématique des Jeux*. Bruselas: Stevens, 1930.

O. NEUGEBAUER: *Vorlesungen über Geschichte der antiken mathematischen Wissenschaften*. Volumen I: *Vorgriechische Mathematik*. Berlín: Springer, 1934.

H. POINCARÉ: *The Foundations of Science*. Lancaster, Pa.: Science Press, 1913.

H. RADEMACHER y O. TOEPLITZ: *Von Zahlen und Figuren*, 2.ª edición. Berlín: Springer, 1933.

B. RUSSELL: *Introducción a la filosofía matemática* (en *Obras escogidas* de B. R.) Madrid. Aguilar, 1956.

—*Los principios de la matemática*. Buenos Aires. Espasa-Calpe, 1948.

D. E. SMITH: *A Source Book in Mathematics*. Nueva York: McGraw-Hill, 1929.

H. STEINHAUS: *Mathematical Snapshots*. Nueva York: Stechert, 1938.

H. WEYL: «The Mathematical Way of Thinking», *Science*, XCII (1940), págs. 437 y sgs.

— *Philosophie der Mathematik und Naturwissenschaft*, Handbuch der Philosophie, Bd. II, Munich: Oldenbourg, 1926, páginas 3-162.

CAPÍTULO I

L. E. DICKSON: *Introduction to the Theory of Numbers*. Chicago: University of Chicago Press, 1931.

— *Modern Elementary Theory of Numbers*. Chicago: University of Chicago Press, 1939.

G. H. HARDY: «An Introduction to the Theory of Numbers», *Bulletin of the American Mathematical Society*, XXXV (1929), págs. 789 y sgs.

G. H. HARDY y E. M. WRIGHT: *An Introduction to the Theory of Numbers*. Oxford: Clarendon Press, 1938.

J. V. USPENSKY y M. H. HEASLET: *Elementary Number Theory*. Nueva York: McGraw-Hill, 1939.

CAPÍTULO II

G. BIRKHOFF y S. MACLANE: *Álgebra moderna*. Barcelona. Teide, 1954.

M. BLACK: *The Nature of Mathematics*. Nueva York: Harcourt, Brace, 1935.

T. DANTZIG: *Number, the Language of Science*, 3.ª edición. Nueva York: Macmillan, 1939..

G. H. HARDY: *A Course of Pure Mathe-*

matics, 7.ª edición. Cambridge: University Press, 1938.

K. Knopp: *Theory and Application of Infinite Series*, traducción de Miss R. C. Young. Londres: Blackie, 1928.

A. Tarski: *Introduction to Logic*. Nueva York: Oxford University Press, 1939.

F. Enriques: *Para la historia de la lógica*. Buenos Aires. Espasa-Calpe, 1953.

CAPÍTULO III

J. L. Coolidge: *A History of Geometrical Methods*. Oxford: Clarendon Press, 1940.

A. De Morgan: *A Budget of Paradoxes*, 2 vols. Chicago: Open Court, 1915.

L. E. Dickson: *New First Course in the Theory of Equations*. Nueva York: Wiley, 1939.

F. Enriques (director de la edición): *Fragen der Elementargeometrie*, 2.ª edición, 2 vols. Leipzig: Teubner, 1923.

E. W. Hobson: «*Squaring the Circle*», *a History of the Problem*. Cambridge: University Press, 1913.

A. B. Kempe: *How to Draw a Straight Line*. Londres: Macmillan, 1877.

F. Klein: *Famous Problems of Geometry*, traducido por W. W. Beman y D. E. Smith, 2.ª edición. Nueva York: Stechert, 1930.

L. Mascheroni: *La geometria del compasso*. Palermo: Reber, 1901.

G. Mohr: *Euclides Danicus*. Copenhague: Holst, 1928.

J. M. Thomas: *Theory of Equations*. Nueva York: MacGraw-Hill, 1938.

L. Weisner: *Introduction to the Theory of Equations*. Nueva York: Wiley, 1939.

CAPÍTULO IV

W. C. Graustein: *Introduction to Higher Geometry*. Nueva York: Macmillan, 1930.

D. Hilbert: *Fundamentos de la geometría*. Madrid. C. S. I. C., 1952.

C. W. O'Hara y D. R. Ward: *An Introduction to Projective Geometry*. Oxford: Clarendon Press, 1937.

G. de B. Robinson: *The Foundations of Geometry*. Toronto: University of Toronto Press, 1940.

Girolamo Saccheri: *Euclides ab omni naevo vindicatus*, traducido por G. B. Halsted. Chicago: Open Court, 1920.

R. G. Sanger: *Synthetic Projective Geometry*. Nueva York: McGraw-Hill, 1939.

O. Veblen y J. W. Young: *Projective Geometry*, 2 vols. Boston: Ginn, 1910 y 1918.

J. W. Young: *Projective Geometry*. Chicago: Open Court, 1930.

CAPÍTULO V

P. Alexandroff: *Einfachste Grundbegriffe der Topologie*. Berlín: Springer, 1932.

D. Hilbert y S. Cohn-Vossen: *Anschauliche Geometrie*. Berlín: Springer, 1932.

M. H. A. Newman: *Elements of the Topology of Plane Sets of Points*. Cambridge: University Press, 1939.

H. Seifert y W. Threlfall: *Lehrbuch der Topologie*. Leipzig: Teubner, 1934.

CAPÍTULO VI

R. Courant: *Differential and Integral Calculus*, traducido por E. J. McShane, edición revisada, 2 vols. Nueva York: Nordemann, 1940.

G. H. Hardy: *A Course of Pure Mathematics*, 7.ª edición. Cambridge: University Press, 1938.

W. L. Ferrar: *A Text-book of Convergence*. Oxford: Clarendon Press, 1938.

Para la teoría de las fracciones continuas, véase:

S. Barnard y J. M. Child: *Advanced Algebra*. Londres: Macmillan, 1939.

CAPÍTULO VII

R. Courant: Soap Film Experiments with Minimal Surfaces», *American Mathematical Monthly*, XLVII (1940), págs. 167 - 74.

J. Plateau: «Sur les figures d'équilibre d'une masse liquide sans pésanteur», *Mémoires de l'Académie Royale de Belgique*, nouvelle série, XXIII (1949)

— *Statique expérimentale et théoretique des Liquides*. París: 1873.

CAPÍTULO VIII

C. B. Boyer: *The Concepts of the Calculus*. Nueva York: Columbia University Press, 1939.

R. Courant: *Differential and Integral Calculus*, traducido por E. J. McShane, edición revisada, 2 vols. Nueva York: Nordemann, 1940.

G. H. Hardy: *A Course of Pure Mathematics*, 7.ª edición. Cambridge: University Press, 1938.

ÍNDICE ALFABÉTICO

ÍNDICE ALFABÉTICO DE MATERIAS